

NORTHWESTERN UNIVERSITY

Evaluative Mindsets and the Influence of False Information

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Cognitive Psychology

By

Nikita A. Salovich

EVANSTON, ILLINOIS

June 2022

© Copyright by Nikita A. Salovich, 2022

All Rights Reserved

Abstract

People are exposed to inaccurate claims and ideas every day, from sources intended to inform, entertain, or do both. A large body of research has demonstrated that exposure to inaccurate statements, even when conveying obviously false ideas, can affect people's subsequent judgments. Contemporary accounts suggest that these effects may be due to people's failure to evaluate information during exposure, increasing the likelihood that false information will be encoded and available for retrieval on subsequent tasks. The goal of this dissertation is to better understand how the task goals and mindsets people adopt when comprehending information affect the likelihood they are influenced by false claims and ideas. In Chapter 1, we demonstrate that deliberately evaluating the accuracy of information can reduce reproductions of inaccuracies, as well as encourage people's use of correct prior knowledge. In Chapter 2, we examine whether people can be encouraged to develop and maintain evaluative mindsets without explicit instruction. Specifically, we test whether confronting people with their susceptibility to inaccurate information can prompt evaluative mindsets, thereby reducing the extent to which they are influenced by false claims. In Chapter 3, we explore how both evaluative and non-evaluative tasks and mindsets might differentially affect the influence of inaccuracies presented on social media. Understanding the circumstances that encourage and discourage evaluation can offer insight into processes underlying people's susceptibility to false information, and can inform the design of real-world applications intended to support a better informed society.

[This page is intentionally blank]

[This page is intentionally blank]

Table of Contents

Abstract	3
Acknowledgments	4
Table of Contents	6
List of Figures	7
Introduction	9
Chapter 1	22
Experiment 1	22
Experiment 2	34
Experiment 3	43
Chapter 2	56
Experiment 4	59
Experiment 5	74
Chapter 3	86
Experiment 6	90
Experiment 7	102
Conclusion	126
References	132
Appendix A	148
Appendix B	149

List of Tables, Illustrations, Figures, and Graphs

Figure 1. <i>Error Rates in Experiment 1</i>	30
Figure 2. <i>Correct Response Rates in Experiment 1</i>	31
Table 1. <i>Mean rates of incorrect lure and correct response production in Experiment 1. Judgment type varied between-subjects only.</i>	32
Figure 3. <i>Error Rates in Experiment 2</i>	39
Figure 4. <i>Correct Response Rates in Experiment 2</i>	40
Table 2. <i>Mean rates of incorrect lure and correct response production in Experiment 2. Judgment type varied within subjects</i>	41
Figure 5. <i>Error Rates in Experiment 3</i>	46
Figure 6. <i>Correct Response Rates in Experiment 3</i>	47
Table 3. <i>Mean rates of incorrect lure and correct response production in Experiment 3. Judgment type varied between and within subjects depending on condition.</i>	48
Figure 7. <i>Visual Depiction of Experiment 4 Procedure</i>	63
Figure 8. <i>Error Rates Following False Information in Experiment 4</i>	67
Figure 9. <i>Correct Response Rates Following False Information in Experiment 4</i>	70
Table 4. <i>Mean Rates of Incorrect Lure and Correct Response Production in Experiment 4</i>	71
Figure 10. <i>Error Rates Following False Information in Experiment 5</i>	77
Figure 11. <i>Correct Response Rates Following False Information in Experiment 5.</i>	79
Table 5. <i>Mean Rates of Incorrect Lure and Correct Response Production in Experiment 5</i>	80
Table 6. <i>Example Statement and Converted Twitter Stimuli</i>	92
Figure 12. <i>Error Rates Following False Information in Experiment 6</i>	96
Figure 13. <i>Correct Response Rates Following False Information in Experiment 6.</i>	98

Table 7. <i>Mean Rates of Incorrect Lure and Correct Response Production in Experiment 6</i>	98
Figure 14. <i>Error Rates Following False Information in Experiment 7</i>	109
Figure 15. <i>Correct Response Rates Following False Information in Experiment 7.</i>	111
Table 8. <i>Mean Rates of Incorrect Lure and Correct Response Production in Experiment 7 . . .</i>	112
Figure 16. <i>Self-Reported Consideration of Accuracy in Experiment 7.</i>	115
Table 9. <i>Mean Rates of Incorrect Lure Reproduction and Correct Response Production Based on Self-Reported Considerations of Accuracy Experiment 7.</i>	117

Evaluative Mindsets and the Influence of False Information

People are exposed to inaccurate claims and ideas every day from sources intended to inform (e.g., news outlets), entertain (e.g., popular fiction), or potentially do both (e.g., Twitter feeds). False and misleading information is a regular component of people's experiences, with concern for the circulation and influence of inaccurate information arguably at an all-time high. These concerns are well-founded, as a growing body of research has documented the routine dissemination of false ideas. Rumors and hoaxes spread rapidly online, with fact-checked inaccurate information spreading further and faster than fact-checked accurate information (Vosoughi et al., 2018). Major world events have even been attributed to the circulation of false information, including the results of the 2016 U.S. election, the 2019 measles outbreak, and hesitancy towards COVID-19 vaccines (Allcott & Gentzkow, 2017; Carrieri et al., 2019; Loomba et al., 2021; Tucker et al., 2018).

Contrary to popular belief, people are not only influenced by false information that seems plausible, or by inaccuracies they are motivated to believe are true. Empirical projects have demonstrated that people's judgments and decisions are adversely affected by information they should already know is false (e.g., Fazio et al., 2013; Fazio, Brashier, et al., 2015; Fazio et al., 2019; Rapp, 2008; Rapp, 2016). This includes inaccurate assertions about real-world phenomena (e.g., Toothbrushing causes gum disease; Gerrig & Prentice, 1991; Prentice et al., 1997), false declarative statements (e.g., Narcolepsy is the term for an inability to sleep; Hinze et al., 2014; Marsh et al., 2003), and news headlines conveying incorrect claims (e.g., "Coconut oil's history in destroying viruses, including Coronaviruses"; Pennycook, Binnendyk, et al., 2021). After exposure to inaccurate ideas, people misperceive them as more true (e.g., on Likert-scales; Brashier et al., 2020; Fazio et al., 2019; Fazio, Brashier, et al., 2015), judge related inaccurate

claims as valid (e.g., via true/false judgments; Rapp, Hinze, et al., 2014; Salovich et al., 2021), and reproduce the inaccuracies to answer related questions (Donovan & Rapp, 2020; Fazio, Dolan, et al., 2015; Marsh et al., 2003).

Evaluation and the Influence of False Information

Contemporary accounts have questioned whether these effects can be attributed to people's failure to think critically about the accuracy of information during comprehension (e.g., Bago et al., 2020; Fazio, 2020; Pennycook, Epstein, et al. 2021; Pennycook & Rand, 2019; Rapp, Hinze, et al., 2014; Salovich & Rapp, 2021). While some research suggests that people can automatically detect inconsistent and incoherent information (Isberner & Richter, 2014b, Richter, 2015; Richter et al., 2009; Singer, 2006, 2013), such spontaneous evaluation may not always occur or resolve effectively. For example, people incorrectly answer questions containing seemingly obvious semantic anomalies (e.g., answering "two" when asked "how many animals of each kind did Moses take on the ark?" despite the Biblical story involving Noah and not Moses; Brashier & Marsh, 2020; Cook et al., 2018; Erickson & Mattson, 1981). These incorrect responses would be unlikely if people detected the seemingly obvious inaccuracies, or if they successfully resolved the anomalies to produce a correct response.

Analogous effects arise when people are asked questions after exposure to declarative inaccuracies. Rather than discounting false or misleading ideas during exposure, people appear to routinely encode inaccurate content as presented. This problematically increases its availability for retrieval on later tasks (e.g., Kelley & Lindsay, 1993) and the chance it is perceived as true when encountered again (e.g., Reber & Unkelbach, 2010; Arkes et al., 1991) despite any correct prior knowledge one may possess (see Brashier & Marsh, 2020 for review). This suggests inaccurate exposures may go unnoticed, and even if detected, can remain influential on

subsequent tasks (Rapp & Salovich, 2018). Any nonstrategic evaluative processes therefore may not effectively overcome the effects of inaccurate exposures.

In contrast to any passive evaluative processes, *deliberate evaluation* involves conscious, goal-driven contemplation of the accuracy of information (Cook & O'Brien, 2014; Isberner & Richter, 2014a; Wiswede et al., 2013). An accumulating body of work suggests that activities that explicitly encourage evaluation, such as instructions to edit text content for accuracy, can reduce the errors associated with prior exposure to inaccuracies. For example, Rapp, Hinze, et al. (2014) presented participants with a text containing well-known accurate and inaccurate information. To encourage evaluation, some participants were asked to “behave like fact-checkers” and correct any false information they read in the text. After reading, participants judged single statements summarizing accurate or inaccurate information from the text (e.g., “Toothbrushing causes gum disease”) as true or false. Participants made more judgment errors after previously reading inaccurate as compared to accurate information, but those errors were reduced when participants were tasked with fact-checking during reading.

Other studies have leveraged a similar fact-checking manipulation to test whether encouraging participants to focus on accuracy can reduce susceptibility to the illusory truth effect, i.e., the finding that repeated exposure to information (including false information) encourages people to believe it is true (e.g., Brashier & Marsh, 2020; Fazio, Brashier, et al., 2015; Hasher et al., 1977; Hassan & Barber, 2021). In one study, Hawkins and Hoch (1992) presented participants with true and false consumption-oriented statements (e.g., “Stone-ground flour retains more nutrients than conventional flour”), with half of the participants asked to rate the accuracy of each statement, and the other half asked to rate how easy the statement was to understand. Afterwards, all participants judged the truth of previously viewed and new

statements. Participants overall rated repeated statements as more true than new statements, but this effect was substantially smaller for participants who made accuracy than easiness judgments at exposure.

In a more recent examination by Brashier et al. (2020), participants initially made evaluative accuracy judgments or non-evaluative interest judgments about true and false general knowledge statements of varying difficulties. Again, deliberate evaluation of the false statements during exposure appeared to overcome repetition-based effects. While participants who previously judged statements for interest showed the illusory truth effect, participants who previously made accuracy judgments showed no difference in truth ratings between repeated and non-repeated statements, specifically when the information was well-known to be false. This highlights the potential role of relevant prior knowledge in exercising and obtaining benefits associated with deliberate evaluation.

How Does Evaluation Protect Against False Information?

While an accuracy focus may help reduce effects associated with exposures to inaccuracies, the mechanism by which this influence might be reduced remains unclear. One possibility is that deliberate evaluation can reduce the extent to which inaccuracies interfere with people's use of their correct understandings. As previously noted, exposure can increase the salience and availability of information in memory, as well as the ease with which it can be retrieved (Kelley & Lindsay, 1993; Rapp, 2016). Some processing accounts explain the influence of obvious falsehoods on people's judgments as due to competition and/or interference between recently encoded information and prior knowledge (Lewis & Anderson, 1976; Rapp, Hinze, et al., 2014; Rapp, Jacobina, et al., 2014; Weil et al., 2020). When people read inaccurate information (e.g., "Narcolepsy is the term for an inability to sleep"), they may encode incorrect

associations between concepts (e.g., “inability to sleep” and “narcolepsy”) without sufficient activation of, or even alongside accurate associations (e.g., “inability to sleep” and “insomnia”). If the inaccurate concept (e.g., “narcolepsy”) is more familiar or shares sufficient semantic associates with background knowledge (e.g., “sleep,” “sleep difficulties,” “medical conditions”), it may retain greater availability than existing accurate understandings (e.g., “insomnia;” Anderson, 1981; Rapp, Jacobina, et al., 2014; Storm, 2011), affording use of the false answer over correct knowledge on later tasks. Exposures to inaccuracies can therefore cause people to become confused about what is actually true in the world, to doubt their accurate knowledge, and to potentially neglect it in favor of recently encountered falsehoods (see Rapp & Salovich, 2018 for a review). Accuracy judgments may disrupt these interference effects by encouraging participants to activate relevant prior knowledge during comprehension and make an informed decision about the validity of presented information. This would increase both the likelihood of detecting false information when it contradicts what is already known to be true, and the availability of accurate prior knowledge to be privileged on subsequent tasks.

The process of monitoring and detecting changes to information content has generally been shown to enhance memory performance, specifically for cases in which interference reduces the retrieval of prior knowledge. Consider cases in which recently presented information interferes with the retrieval of previously encoded information about an event (e.g., Belli, 1989; also see Otero & Kintsch, 1992 for related findings). Prompts to monitor text and event descriptions have been shown to reduce such retroactive interference effects in paired-associates word learning tasks (Negley et al., 2018) and misinformation paradigms (Loftus, 1979). Researchers have interpreted these findings through a *memory-for-change* account (e.g., Negley et al., 2018; Wahlheim, 2015; Wahlheim & Jacoby, 2013) based on earlier findings on recursive

reminders (e.g., Hintzman, 2004). According to this account, “noticing contradiction requires that the earlier-read claim be brought to mind by reading the later-made claim, which serves as an implicit repetition of the earlier-made claim that could increase the probability of its later recall” (Negley et al., 2018, p. 1). When people detect a discrepancy, earlier information is retrieved from memory and encoded into the representation of the subsequent inconsistent information. This increases the availability and successful recall of that earlier information over the more recently encountered misinformation, suggesting routine operations of memory can lead to and combat interference effects.

While this logic has been applied to explain episodic retrieval (e.g., involving word lists or unfolding event details presented in an experimental context), there is reason to believe it may apply to a more general set of comprehension experiences. For example, work on text comprehension has highlighted the importance of detection and evaluation via the activation of background knowledge (e.g., Cook & O’Brien, 2014; O’Brien & Cook, 2016a; O’Brien & Cook, 2016b; Richter et al., 2009; Richter, 2015). According to the *Resonance-Integration-Validation* (RI-Val) model of comprehension (Cook & O’Brien, 2014; O’Brien & Cook, 2016b), newly encoded information and related concepts from long-term memory are activated through a passive, resonance-like mechanism (resonance stage), and then integrated or “linked” in active memory (integration stage). These linkages are subsequently validated or compared against information in long-term memory (validation stage). As such, detecting “mismatches” between encoded information and existing knowledge, which is crucial for successful learning and comprehension, depends on the salience and availability of relevant background knowledge (Richter, 2015; Singer, 2019).

The above accounts are leveraged here to propose that accuracy judgments may prompt retrieval of accurate prior knowledge that facilitates the evaluation of presented information and the detection of inaccuracies. This would afford concurrent activation and encoding of correct knowledge alongside any inaccuracy. While falsehoods may still be encoded and integrated during comprehension given the routine operations of memory, any such traces would also include corrections to the information (Rapp, Hinze, et al., 2014). Both memory-for-change and RI-Val frameworks predict that evaluation increases the availability of correct prior knowledge so as to outcompete inaccurate encodings. We therefore expect that accuracy prompts should not only reduce reliance on blatantly false information, but also that subsequent responses should demonstrate greater reliance on correct prior knowledge.

Evaluation and Reliance on Correct Prior Knowledge

Previous work has often examined the efficacy of evaluation using forced choice, closed-ended tasks including binary true/false judgments (e.g., Rapp, Hinze, et al., 2014) and ratings on Likert-scales (e.g., Brashier et al., 2020; but see Marsh & Fazio, 2006). These measures do not permit separate examination of errors and correct responses, as the selection of one response (e.g., “false”) entails the negation of the other (e.g., not “true”). In the present study we collected people’s answers via open-ended questions, which allows for a diversity of participant responses including correct answers, incorrect answers *other than* any presented false ideas (e.g., incorrect responses that participants may not have read) or withheld responses reflecting participant uncertainty (Rapp & Salovich, 2018; Salovich et al., 2021). This affords a more nuanced understanding of the downstream consequences of evaluation. Importantly, it also provides an opportunity to test the prediction that the benefits associated with an accuracy focus can be attributed to increased availability of correct prior knowledge.

This is a theoretically important question, as accuracy judgments could reduce people's endorsements of false claims without necessarily encouraging participants to rely on their existing understandings in lieu of the inaccuracies. Some work even explicitly suggests that people do not need to assess presented information with prior knowledge to determine whether a sentence is false. In recent work, Weil and Mudrik (2020) asked participants to read statements, some of which were false (e.g., "Macadamia is a type of berry"). Having participants activate correct prior knowledge immediately preceding presentations of inaccuracies did not facilitate performance when compared to instances in which prior knowledge was not explicitly invoked. The researchers concluded that "the detection of falsehood does not require bringing the truth to mind" (p. 13), but rather that people may evaluate statements by detecting semantic mismatches in sentence content (e.g., between "macadamia", a type of nut, and "berries", a type of fruit; also see Rapp, 2008.) Accuracy judgments could therefore reduce reliance on inaccuracies *without* necessarily increasing correct responses. This also begs the question whether reductions in the influence of inaccurate information could be attributed to people becoming overall more conservative in what they report as true, rather than their evaluations specifically affording reliance on correct understandings. Consider that warnings intended to reduce reliance on inaccuracies have at times resulted in reductions to *both* incorrect and correct responses (Andrews-Todd et al., 2021; Marsh & Fazio, 2006; Salovich et al., 2021). Participants therefore could be more inclined to withhold answers or respond that they are uncertain following accuracy judgments rather than making use of their correct prior knowledge.

The experiments comprising this dissertation tested whether evaluation influences *both* correct and incorrect responses, providing a crucial assessment of the mechanistic underpinnings of potential evaluative benefits. As increasing attention is placed on evaluation as a promising

method for supporting intentional and informed interactions with information (e.g., Pennycook & Rand, 2021), understanding the mechanisms that support effective evaluation and their consequences proves of theoretical and practical importance. The current dissertation represents a necessary examination for adjudicating between the above-mentioned theoretical possibilities.

Establishing and Sustaining Evaluative Mindsets

An additional issue considered in this dissertation involves determining what is required to establish and maintain an evaluative focus during comprehension. Explicit instructions to detect and correct inaccuracies have obtained consistent benefits for establishing an accuracy focus (Brashier et al. 2020; Calvillo & Smelter, 2021; Rapp, Hinze, et al., 2014). But the benefits of this focus remain uncertain with respect to whether any evaluative tendency is maintained when the requirement to make accuracy judgments is removed or only intermittently prompted. It is also uncertain whether accuracy benefits are restricted to information people are explicitly asked to evaluate, or if benefits might also extend to other content presented in the same experience (including when that information is associated with a non-evaluative task or judgment).

These considerations crucially underlie contemporary accounts of so-called evaluative mindsets, involving recurring “modes of thought” that support validation and comprehension (e.g., Mayo, 2015, 2019). Research suggests that increased levels of skepticism can be triggered by factors such as the qualities of a particular source (Andrews & Rapp, 2014; Schul et al., 2004; Sparks & Rapp, 2011), contextual features of a task (e.g., Mayo et al., 2014; Wiswede et al., 2013), or due to individual differences in the propensity to trust (or distrust) people or information (Mayo et al., 2014; Mayo, 2015). Recent work has examined whether pre-exposure tasks can “prime” an accuracy focus that extends to subsequently presented information

(Pennycook, Epstein, et al., 2021; Pennycook et al., 2020; Salovich & Rapp, 2021). However, some of these studies have reported small effects with difficult-to-replicate results (e.g., Roozenbeek et al., 2021). In Chapter 1 of this dissertation, we examined whether intermittent accuracy judgments encourage people to adopt a general accuracy focus that is maintained even when subsequent information is presented without an evaluative prompt. We expanded on this idea in Chapter 2, investigating whether evaluative mindsets can be encouraged without explicit instruction to engage in evaluation. Specifically, we tested whether and how confronting people about their potential susceptibility to the influence of false information might motivate evaluation. Given routine demonstrations that people fail to critically engage with information when left to their own devices, evaluations may be contingent on explicit prompting, reflecting in-the-moment considerations rather than overarching evaluative mindsets.

Prior work has assessed evaluative mindsets through people's in-the-moment processing of information using response latencies and EEG activity (e.g., Hagoort et al., 2004; Isberner & Richter, 2014b; Wiswede et al., 2013). Others have considered the downstream consequences of evaluative processing, such as whether people's judgments and decisions are suggestive of increased deliberation or skepticism toward information (e.g., Andrews & Rapp, 2014; Mayo, 2019; Sparks & Rapp, 2011). In this dissertation, we assess the degree to which an evaluative mindset is established and sustained through people's uptake of and reliance on presented inaccuracies. If the manipulations in question successfully encourage the routine evaluation of information, then responses following exposure to inaccuracies should resemble responses in conditions requiring explicit evaluation, which traditionally has been characterized with lower rates of reliance on falsehoods (and as we find in Chapter 1, also higher rates of correct responses). In Chapter 3, we also begin to consider other in-the-moment measures that may

capture evaluative strategies (e.g., response latencies), in addition to people's qualitative self-reported considerations of accuracy when engaging with presented information.

Overview of Dissertation

The goal of the proposed dissertation is to understand how the tasks and mindsets with which people approach information affect the likelihood they are influenced by false claims and ideas. It follows a multiple manuscript format and includes three papers. Specifically, it consists of three empirical chapters (each representing a separate study) resulting in a total of seven experiments.

Chapter 1 investigates whether deliberately evaluating the accuracy of information can reduce reproductions of inaccuracies as well as encourage people's use of their correct prior knowledge. It includes a set of three experiments published as an original empirical manuscript in *Cognition* (Salovich, Kirsch, et al., 2022). In Experiment 1, participants who were instructed to engage in deliberate evaluation of potentially inaccurate statements reproduced fewer inaccurate ideas and produced more correct answers to post-reading questions than did participants who simply rated their interest in the inaccurate statements. In Experiments 2 and 3, the same benefits were obtained even when participants were not consistently prompted to evaluate the statements. These results offer insight into how and when evaluation can encourage participants to rely on correct prior knowledge over presented inaccuracies, as well as what is required to establish and maintain an evaluative mindset.

Chapter 2 describes two experiments (currently under review at *Journal of Applied Research in Memory and Cognition*) that take the evaluative mindset account a step further, testing whether people can be encouraged to adopt an accuracy focus during comprehension without explicit instruction to do so. We examined whether and how confronting people about

their potential susceptibility to the influence of false information might motivate evaluation and reduce the problematic effects associated with exposures to inaccurate ideas. Participants made non-evaluative interest ratings about true and false statements, and then answered related general knowledge questions. In Experiment 4, participants who received positive or negative performance feedback about their susceptibility to inaccurate information reproduced fewer incorrect ideas and produced more correct answers than did participants who did not receive feedback. In Experiment 5, analogous benefits emerged when participants were simply informed that their use of false information was being monitored. The experimental findings provide further support that people's thoughts and beliefs about their own resistance to inaccuracies play a crucial role in how they approach and are subsequently affected by false and misleading claims. The results also have practical implications for reducing people's belief in and the spread of false and misleading information.

Chapter 3 includes two experiments (currently in-prep for publication) that explore how the tasks and mindsets that participants adopt during exposures to social media content might differentially affect reliance on inaccuracies. Increasing concerns about the belief and spread of false information on social media (e.g., Lewandowsky et al., 2017) make it an opportune time to examine whether challenges and benefits highlighted in prior experiments might emerge in naturalistic information environments. We modified the general knowledge statements used in the experiments reported in Chapters 1 and 2 into realistic tweets that contained either correct or incorrect information. In Experiment 6, we found that exposures to false information conveyed in tweets increased the rate at which those inaccuracies were reproduced by participants to answer general knowledge questions, as compared to after participants viewed true information or no relevant information. Exposures to false tweets also problematically reduced participants'

correct answers relative to after viewing true tweets or no information. In Experiment 7, we explored how interest, accuracy, and like judgments affected participants' susceptibility to inaccuracies presented in a social media context. Participants who judged tweets for accuracy at exposure were less likely to reproduce inaccuracies from tweets to answer subsequent questions, and more likely to answer with their accurate prior knowledge, than were participants who rated how interesting the tweets were, or participants who rated how likely they would be to "like" the tweet on social media. Analyses of response times, relationships between ratings of interest, accuracy, and "liking" a tweet with participants' inaccurate reproductions, and participants' qualitative testimonies concerning their accuracy contemplations corroborated the unique importance of accuracy judgments for reducing the effects of inaccurate exposures. Chapter 3 offers practical implications to help combat the belief in and spread of misinformation online, which is important given concerns about the amount of false and misleading information available on social media.

Chapter 1:

Evaluative Mindsets Can Protect Against the Influence of False Information

Across three experiments, we examined how and when evaluation reduces the influence of inaccurate information and supports reliance on accurate knowledge. In Experiment 1, participants rated true and false statements on either interest or accuracy, and were subsequently asked a series of open-ended questions related to the contents of those statements. This allowed us to test the utility of evaluation for overcoming exposure to false information, and whether evaluations support people's correct responses derived from prior knowledge. Experiments 2 and 3 examined whether the evaluative focus elicited by accuracy judgments would be maintained when intermittently prompted by the task, and any resulting effects on incorrect and correct responses. The findings provide insight into when and why evaluation can overcome the effects of inaccurate exposures, the degree to which evaluative mindsets are amenable to prompting, and whether any downstream benefits are sustained when evaluation is not consistently encouraged or expected.

Experiment 1

We investigated the effects of deliberate evaluation on people's reproductions of false information and on their accurate productions. Participants read a series of true statements (e.g., "The largest planet in the solar system is Jupiter"), false statements (e.g., "The largest planet in the solar system is Saturn"), and semantically-related "filler" statements (e.g., "The brightest star in the night sky is Sirius"). Half of the statements were easy, meaning they were likely well-known to participants (e.g., "The capital of France is Paris/Marseille") and half were hard, meaning they were less likely to be known to participants (e.g., "The U.S. Naval Academy is located in Annapolis/Washington"). Participants were randomly assigned to rate the statements

on either how interesting each was or how accurate each was. Afterwards they were asked open-ended questions corresponding to each of the previously presented statements (e.g., “What is the largest planet in the solar system?”). We were specifically interested in productions of inaccuracies from the statements (referred to as *incorrect lures*; e.g., “Saturn”) and correct responses (e.g., “Jupiter”).

In line with prior work (Donovan & Rapp, 2020; Eslick et al., 2011; Fazio, Dolan, et al., 2015; Marsh & Fazio, 2006; Marsh et al., 2003), we expected participants would answer questions with an incorrect lure more often after exposure to statements containing those lures as compared to after reading true statements. We also expected incorrect lure reproductions to happen more often after exposure to lesser known, hard statements as compared to well-known, easy statements. Additionally, participants should produce fewer correct responses after exposure to incorrect lures as compared to after exposure to true statements, particularly when the statements are easy versus hard.

We also hypothesized that the type of judgment that participants were asked to make while reading the statements would influence rates of inaccurate reproductions and correct productions. First, we predicted that participants tasked with judging the accuracy of statements would be *less* likely to reproduce incorrect lures than would participants asked to rate how interesting the statements were. We predicted that this benefit would be more readily apparent for easy as compared to hard statements, as participants could leverage accurate knowledge to inform their accuracy judgments for more familiar as compared to less familiar items. Second, we hypothesized that participants tasked with making accuracy judgments would be more likely to provide correct answers than would participants tasked with making interest judgments. Again, we expected this to occur more so for easy as compared to hard statements as participants

could rely on their relevant and accurate prior knowledge about familiar ideas despite the incorrect information they had read.

Method

Participants

Given our plan to analyze the repeated-measures data with mixed effect models, we aimed to recruit approximately 80 participants per between subject condition (i.e., 160 participants for Experiment 1). This allowed for sufficient observations of each item type both within and across participants to achieve adequate power while reducing Type 1 Error rates (see Luke, 2017). We therefore recruited one hundred and eighty Amazon Mechanical Turk workers to participate in the study, overrecruiting slightly to accommodate potential exclusions. After excluding participants who failed a comprehension check, an English language check, or reported looking up answers, we were left with a sample of 171 participants (89 female; M age = 38.60 years), $n = 92$ in the interest judgment condition and $n = 79$ in the accuracy judgment condition. Across participants, we recorded 13,851 observations. All participants were paid equal to or above the U.S. minimum wage at the time of data collection.

Design

The experiment followed a 2 (judgment type: interest or accuracy) x 3 (statement validity: true, false, or filler) x 2 (statement difficulty: easy or hard) mixed design. Judgment type varied between groups, while statement validity and difficulty were manipulated within-subjects.

Materials

We selected 81 facts from previously published general knowledge norms (Tauber et al., 2013). Forty test items were easy (known by most people) and 41 were hard (known by few people), based on the norms. Each fact included a correct answer and an incorrect lure. Incorrect

lures referred to plausible but incorrect alternatives (e.g., Dried grapes are called “prunes” rather than “raisins”), which were chosen based on commission errors provided during norming. We also generated filler statements semantically related to each of the 81 facts (half true, half false) to use as a control in a baseline comparison condition. This allowed us to assess rates of incorrect lure productions and correct answer productions as compared to when participants were not previously exposed to true or false information.¹ Each fact had three possible statement versions: true (e.g., “The largest planet in the solar system is Jupiter”), false/inaccurate (“The largest planet in the solar system is Saturn”), and filler (e.g., “The brightest star in the night sky is Sirius”). All statement versions were matched for word length. At test, all participants responded to open-ended questions related to the 81 statements (e.g., “What is the largest planet in the solar system?”).

The full set of materials can be found on OSF (osf.io/eaxf6/).

Procedure

After providing informed consent, participants completed the *exposure phase*. They were presented the following instructions: “In the first part of the study, you will read a series of statements one at a time and indicate how [interesting/accurate] you find each statement to be. Some of the statements are true and some of them are false. Later in the study, you will be asked a series of general knowledge questions.” Then, depending on judgment condition, participants

¹ The purpose of including filler items was twofold: First, we wanted to validate assumptions about the difficulty of the facts/statements used in this project, as they have implications for hypothesized performance differences for easy and hard items. For example, we wanted to ensure that participants were more likely to correctly answer easy versus hard questions when they had not been exposed to related information. This would support the claim that participants were more likely to possess prior knowledge for easy as compared to hard items, which is a crucial assumption underlying the prediction that accuracy judgments should differentially affect responses to easy and hard items. Second, including filler items in the exposure phase resulted in statements that participants would read about but not be tested on, helping to reduce the likelihood that participants would interpret the subsequent general knowledge questionnaire as a memory test.

either rated each of the 81 statements for how interesting they were, on a scale from 1 (*very uninteresting*) to 6 (*very interesting*), or rated each of the 81 statements for how accurate they were, on a scale from 1 (*definitely false*) to 6 (*definitely true*). Statements were presented one-at-a-time in a unique random order for each participant. Each participant read 40 easy and 41 hard statements, with one third of the statements true, one third false, and one third filler.²

Immediately after completing the exposure phase, participants completed the *test phase*. They were asked to complete a general knowledge questionnaire and to “answer based on what you believe to be true about the world.” They answered 81 open-ended questions related to the earlier read statements, presented in a unique random order for each participant, by typing their answers in a textbox. They were instructed not to look up any answers, and that they could leave the box blank or write unsure/no answer if they did not know the answer. Participants took approximately 40 minutes to complete the entire experiment.

Results

Questionnaire coding

Participant responses on the questionnaire ($N = 13,851$) were coded as correct (e.g., “Jupiter”), incorrect lure (e.g., “Saturn”), incorrect other (e.g., responding with anything else such as “Pluto” or “Sun”), or unsure/blank. Two raters independently coded a quarter of the responses in the data set, with the remaining coded by one rater only. Inter-rater reliability for dual-coded responses was reliably high ($\kappa = .97$) with all disagreements resolved through discussion. While this paper focuses on rates of incorrect lures and correct responses, all coded data are publicly available on OSF (osf.io/eaxf6/).

² The combination of 40 easy statements and 41 hard statements resulted in a total of 81 statements, which is divisible by three. This ensured equal presentations of true, false, and filler statements in Experiment 1, despite there being slightly uneven distributions of true/false/fillers within easy and hard items.

Models

Data analyses were run using generalized linear mixed effect models (GLMM) with the R packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2014). These mixed effect models simultaneously account for the variance introduced due to overall differences across participants and items (random intercepts), as well as the possibility that manipulations of statement difficulty and accuracy could affect each participant differently (random slopes). By controlling for these random effects, we can more clearly and confidently ascertain the effect of the independent variables on incorrect lure and correct answer productions. Additionally, it better allows us to generalize beyond the sampled participants and items included in the current analysis (e.g., other subjects or stimuli; Baayen et al., 2008). Model specifications and outputs for Experiment 1 are publicly available at <https://rpubs.com/nsalovich/evaluation-exp1>.

We also conducted separate analyses of incorrect lure and correct answer responses to test our hypothesized effects, which is consistent with previous research examining how presentations of information affect people's responses to open-answer questions (e.g., Donovan & Rapp, 2020; Kelley & Lindsay, 1993; Fazio et al., 2013; Fazio & Marsh, 2008; Marsh & Fazio, 2006; Marsh et al., 2003). While the analyses of incorrect and correct responses are likely not independent (i.e., an increase in one response can lead to a decrease in the other, or vice versa), both are presented for thoroughness and ease of comparison with prior work. Tables summarizing the distribution of response types for all three experiments are available as supplemental material on OSF (osf.io/eaxf6/).

Responses to Filler Items

Participants' responses following exposure to filler items were examined to help verify the validity of the materials (see Table 1). First, we verified that participants were more likely to

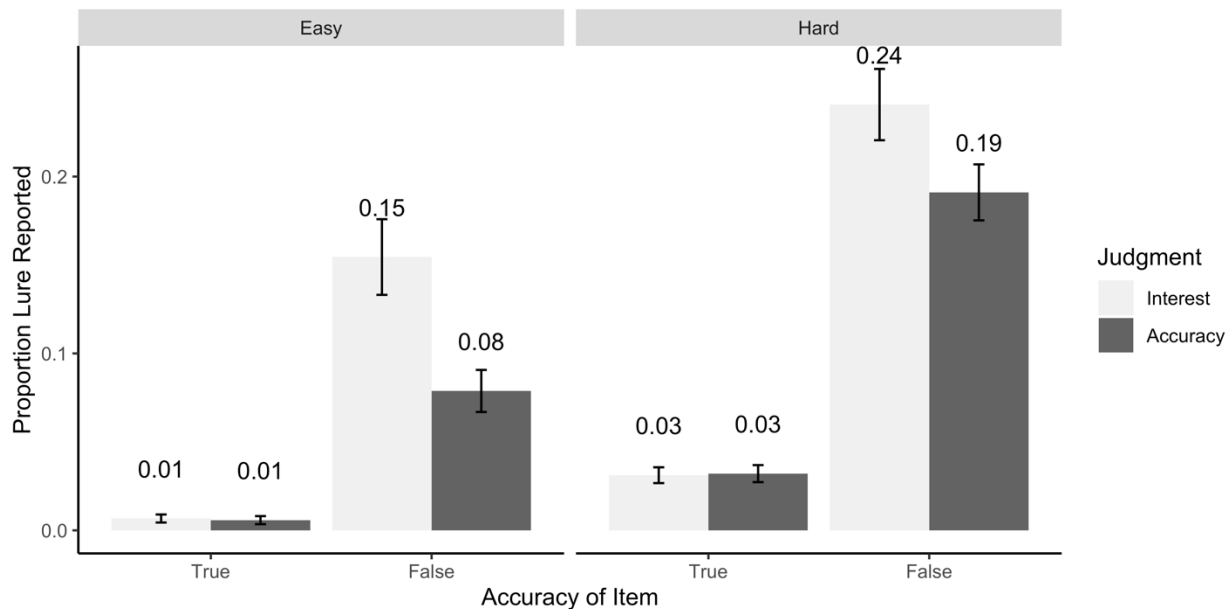
correctly answer questions about easy than hard facts. This would support the claim that participants more likely possessed prior knowledge for easy than for hard facts. We ran two GLMMs; the first predicted *correct answers* following filler statements (correct = 1, all other responses = 0) with difficulty as the single fixed effect (dummy coded; easy = 1, hard = 0), both participants and items as random intercepts, and difficulty as a random slope. As expected, participants produced more correct answers for easy ($M = .79$, $SD = .23$) than hard items ($M = .29$, $SD = .25$), $b = 3.39$, $z = 16.69$ $p < .001$. The second GLMM included the same fixed and random effects, but predicted the rate with which participants spontaneously answered with the *incorrect lure* after viewing filler statements (incorrect lure responses = 1). Participants produced the incorrect lure (e.g., answering “Saturn” to “What is the largest planet in the solar system?”) for only 3.1% of answers, with similar rates for both easy and hard statements, $p > .05$. This suggests that inaccurate responses, if produced, are unlikely to be due to participants’ prior knowledge, but due to exposures during the experiment. All subsequent analyses exclude filler statements, exclusively considering participants’ responses following true and false statements.

Incorrect Lure Productions

Incorrect lure reproduction rates are summarized in Table 1 and Figure 1. To analyze these reproductions, incorrect lure responses were coded as 1 and all other responses as 0. We first examined how statement validity affected lure reproduction. The GLMM included statement validity (contrast coded; true = -.5, false = .5) and statement difficulty (contrast coded; easy = -.5 and hard = .5) as fixed effects, participant and item as random intercepts, and statement validity and difficulty as random slopes. This model simultaneously accounted for variance due to random selection of participants and items, as well as variance in how statement validity and difficulty affected each participant (Richter, 2006).

We observed a main effect of difficulty, with participants reproducing incorrect lures more for hard ($M = .13$, $SD = .16$) than easy questions ($M = .06$, $SD = .13$), $b = 1.67$, $z = 5.06$, $p < .001$. Participants also reproduced more incorrect lures after reading statements containing those inaccuracies ($M = .17$, $SD = .18$) as compared to spontaneously producing those incorrect lures after reading true statements ($M = .02$, $SD = .04$), $b = 2.60$, $z = 13.91$, $p < .001$. As predicted, we also observed a significant statement validity x statement difficulty interaction, $b = -.93$, $z = -2.77$, $p = .006$. To investigate this interaction, simple contrasts were calculated using the emmeans R package (Lenth, 2019). While participants were overall more likely to produce incorrect lures after exposure to false as compared to true statements, this pattern was more readily observed for hard, $z = 13.30$, $p < .001$, than easy questions, $z = 9.67$, $p < .001$. In other words, people were more influenced by lesser-known than well-known inaccuracies.

Did being asked at exposure to judge statements for accuracy as compared to interest affect inaccurate reproductions? To address this question, we ran a GLMM predicting incorrect lure production following exposure specifically to false statements. Fixed effects were statement difficulty (contrast coded; easy = $-.5$ and hard = $.5$) and judgment type (contrast coded; accuracy = $-.5$ and interest = $.5$). Participants and items were again included as random intercepts, and item difficulty and judgment type as random slopes. Overall, participants reproduced fewer incorrect lures if they had earlier made accuracy judgments ($M = .13$, $SD = .14$) as compared to interest judgments ($M = .20$, $SD = .21$), $b = .56$, $z = 2.52$, $p = .01$. The predicted difficulty x judgment type interaction emerged, $b = -.52$, $z = -1.97$, $p = .048$. Simple contrasts revealed that accuracy judgments reduced inaccurate reproductions for easy, $z = 2.71$, $p = .007$, but not hard items, $z = 1.47$, $p = .14$. Evaluating the accuracy of information at exposure reduced incorrect reproductions only for statements for which participants were likely to possess prior knowledge.

Figure 1*Error Rates in Experiment 1*

Note: Error rates at test split by validity of statement at exposure (true or false), difficulty of statement at exposure (easy or hard), and judgment type (interest or accuracy). Judgment type (interest or accuracy) was manipulated between subjects. Error bars represent standard error.

Correct Productions

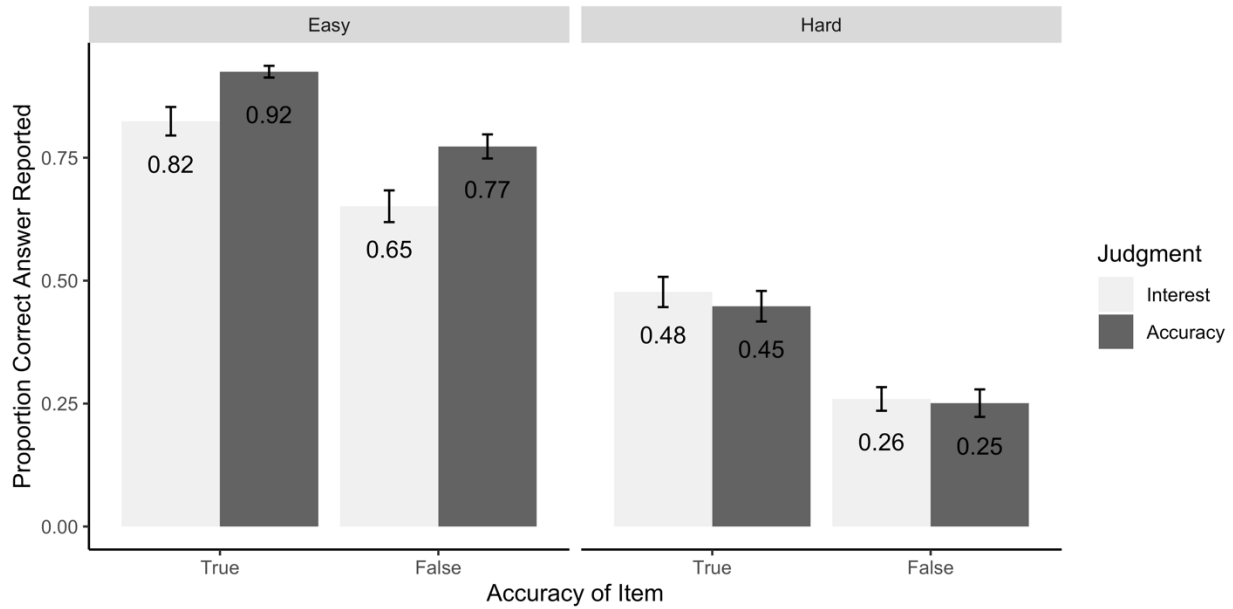
Correct response rates are summarized in Table 1 and Figure 2. To analyze correct answers, responses that were correct were coded as 1 and all other responses as 0. We first examined whether statement validity and difficulty affected correct responses. The GLMM included statement validity (contrast coded; true = .5, false = -.5) and difficulty (contrast coded; easy = .5, hard = -.5) as fixed effects, participant and item as random intercepts, and statement validity and difficulty as random slopes. As expected, participants were more likely to produce correct responses to easy ($M = .79$, $SD = .40$) than hard questions ($M = .36$, $SD = .48$), $b = 3.27$, z

= 15.29, $p < .001$. They also produced more correct responses after exposure to true ($M = .66$ $SD = .47$) as compared to false statements ($M = .48$, $SD = .50$), $b = 1.62$, $z = 14.48$, $p < .001$. There was also a significant statement validity x statement difficulty interaction, $b = .52$, $z = 3.54$, $p < .001$. Simple contrasts revealed that participants were more likely to produce correct answers after exposure to true than to false statements when the statement contents were familiar, $z = 12.83$, $p < .001$, versus unfamiliar, $z = 11.32$, $p < .001$.

We next investigated the effects of accuracy and interest judgments on correct responses. We ran a GLMM predicting correct responses following exposure to false statements as a function of statement difficulty (contrast coded; easy = .5 and hard = -.5) and judgment type (contrast coded; accuracy = .5 and interest = -.5). Participants and items were again included as random intercepts, and item difficulty and judgment type as random slopes. Although there was no main effect of judgment type, $p > .05$, we observed a significant statement difficulty x judgment type interaction, $b = .89$, $z = 3.04$, $p = .002$. Tests of simple contrasts revealed that participants produced more correct responses if they earlier made accuracy as compared to interest judgments for easy statements, $z = 2.95$, $p = .003$, but that pattern was not observed for hard statements, $z = -.34$, $p = .74$. Evaluating false statements increased correct responses, but only when the information was likely to contradict participants' accurate prior knowledge.

Figure 2

Correct Response Rates in Experiment 1



Note: Correct response rates at test split by validity of statement at exposure (true or false), difficulty of statement at exposure (easy or hard), and judgment type (interest or accuracy). Judgment type (interest or accuracy) was manipulated between subjects. Error bars represent standard error.

Table 1

Mean rates of incorrect lure and correct response production in Experiment 1. Judgment type varied between-subjects only.

Statement Validity	Statement Difficulty	Lure	Correct
<i>Interest Instructions</i>			
True	Easy	.01 (.02)	.82 (.28)
	Hard	.03 (.04)	.47 (.29)
False	Easy	.15 (.21)	.65 (.31)
	Hard	.24 (.19)	.26 (.23)
Filler	Easy	.01 (.03)	.76 (.26)
	Hard	.04 (.07)	.28 (.24)
<i>Accuracy Instructions</i>			
True	Easy	.01 (.02)	.92 (.11)

	Hard	.03 (.04)	.45 (.28)
False	Easy	.08 (.11)	.78 (.22)
	Hard	.19 (.14)	.25 (.25)
Filler	Easy	.02 (.04)	.82 (.20)
	Hard	.05 (.07)	.29 (.25)

Note: Numbers in parentheses are standard deviations.

Discussion

A single exposure to true and false statements influenced participants' responses to related questions. Participants were more likely to answer questions with incorrect lures after exposure to those lures than after exposure to true statements, particularly when the statements were hard and less familiar. Participants were also more likely to produce correct answers after exposure to true than to false statements, especially when the statement contents were familiar. While previous work has established that reading true and false statements embedded in texts (e.g., fictional stories) can affect people's subsequent answers to open-response questions (e.g., Donovan & Rapp, 2020; Fazio & Marsh, 2008; Marsh & Fazio, 2006; Marsh et al., 2003), to our knowledge, only one other study to date has reported effects of a similar magnitude for isolated statements (Fazio, Dolan, et al., 2015). This is noteworthy as some accounts suggest people might scrutinize the accuracy of information presented in isolation more so than when embedded in a narrative (e.g., Singer, 2019), where there may be motivation to suspend disbelief to maintain coherence of the unfolding story (Marsh et al., 2003; Marsh & Fazio, 2006; Prentice & Gerrig, 1999). The results here suggest that any such scrutiny did not eliminate effects of exposures to inaccuracies.

The results also demonstrate clear benefits of engaging in deliberate evaluation at exposure. First, being asked to consider the accuracy of statements reduced the influence of false information, as reproductions of incorrect answers were less likely when participants made accuracy as compared to interest judgments. This benefit was observed for easy statements but

not hard statements, which was expected given the predicted mechanism outlined earlier. That is, when statements were easy, accuracy judgments likely afforded participants the opportunity to consider their correct prior knowledge to identify the lures as false. This was not possible for hard statements as they were pre-tested to be unfamiliar to participants and the correct answer likely unknown to them. As further support for this possibility, participants tasked with accuracy judgments produced more *correct* answers than did participants tasked with interest judgments. Also as predicted, accuracy judgments were beneficial at encouraging correct responses to easy but not hard questions, presumably again because relevant prior knowledge was available for retrieval. In sum, evaluations did not lead to overall reductions in responses. Instead, the findings support the hypothesis that deliberate evaluations protect against the influence of false information specifically by increasing the availability and use of correct prior knowledge.

These results replicate and extend previous work demonstrating the utility of deliberate accuracy evaluations. However, it remains less clear whether these benefits are item-specific or reflective of an overall task mindset. If participants were not consistently required to make accuracy judgments, would any benefits associated with an accuracy focus “spill over” to statements that did not explicitly require evaluation? We tested this in Experiment 2.

Experiment 2

In Experiment 2, we examined whether the benefits associated with an explicit accuracy focus would extend to other information in the same experience that did not require considerations of accuracy. Specifically, we tested whether accuracy judgments would prompt an evaluative mindset even when not consistently required, and when other information was associated with a non-evaluative interest judgment. Non-evaluative tasks traditionally require lower levels of involvement and processing during comprehension (Hoch and Hawkins, 1992;

Hawkins et al., 2001), and may even discourage considerations of relevant prior knowledge as it is not explicitly required to complete task goals (e.g., judging information based on comprehensibility, interest, orthographic qualities, etc.; O'Brien & Cook, 2016a). Accuracy judgments may encourage activation of prior knowledge in-the-moment to evaluate specific statements or ideas (i.e., item-specific benefits), but whether they can prompt an overarching mindset sustained during non-evaluative tasks remains unclear.

Some research suggests this spillover may be possible. For example, participants asked to initially rate the accuracy of a single headline were three-times less likely to consider sharing false headlines than were participants who were not asked to evaluate the single headline (Pennycook et al., 2020). In other work, participants given a brief, accuracy prompt were less likely to subsequently share misleading content on social media than were people who did not receive an accuracy “nudge” (Pennycook, Epstein, et al., 2021). These studies suggest an initial accuracy judgment can support participants’ later sharing decisions. Of course, participants’ intentions to *share* information can differ from their actual *belief* in the information. But recent findings by Salovich and Rapp (2021) suggest such benefits may also extend to people’s judgments of truth. In their study, participants who were instructed to recall a time when they were influenced by inaccurate information prior to being exposed to false claims subsequently demonstrated *less* reliance on the claims than did participants who were not asked to recall such an experience. This suggests that tasks that encourage an accuracy focus can help reduce the influence of false information generally, and not just responses to particular claims or statements.

Other recent work, however, suggests that an evaluative mindset may be difficult to motivate and maintain. For example, a recent pre-registered study failed to replicate the benefits of a pre-exposure accuracy nudge for decreasing people’s intentions to share false content

(Roozenbeek et al., 2021). The authors concluded that if an effect exists it may be quite smaller (about 50%) than previously reported, and that the benefit fades over time. They thus called for research to examine what it takes to establish and maintain an accuracy focus, including “further theorization into the mechanisms” underlying any effects, and investigation “into the decay of accuracy nudges” (p. 1176). This aligns with the goal of Experiment 2, which is to understand whether accuracy judgments can encourage an overarching evaluative mindset that extends beyond the information for which evaluations are specifically requested. We also examined the effects of intermittent prompting rather than a single prompt to help determine what is needed to maintain an accuracy focus over repeated presentations of information. This proves especially timely given that recent findings indicate that pairing non-evaluative judgments with evaluative instructions can decrease people’s ability to differentiate true from fake news (Epstein et al., 2022). The scope of evaluative benefits when accuracy and interest judgments are intermixed thus remains a crucial issue to be examined in this experiment.

To address these issues, Experiment 2 tested whether accuracy judgments could exert benefits for information associated with a non-evaluative judgment. We manipulated participants’ requirements to judge each statement for interest or accuracy at exposure *within* rather than *between* subjects. They therefore made both interest and evaluation judgments, with different potential judgments for each statement they read. If evaluative benefits are specific to particular statements that require accuracy judgments, then we expected to again see benefits following accuracy but not interest judgments. However, if accuracy judgments encourage evaluative mindsets that are enacted more broadly, then we expected to see benefits not just for items that required accuracy judgments, but also for items that required interest decisions.

Method

Participants

Eighty³ Amazon Mechanical Turk workers were recruited, none of whom had completed Experiment 1. Seventy-eight (29 female; M age = 38.91 years) remained in the sample after removing participants who failed the comprehension and English language check or reported looking up answers during the study, totaling 6,240 observations across participants. All participants were paid equal to or above the U.S. minimum wage at the time of data collection.

Design

The experiment followed a 2 (statement validity: true or false) x 2 (statement difficulty: easy or hard) x 2 (judgment type: interest or accuracy) within-subjects design. All factors varied within-subjects including judgment type, as all participants now made both interest and accuracy judgments albeit with only one judgment type made for each statement.

Materials

We used 80 of the 81 facts from Experiment 1, eliminating one hard statement (Item 81, “The river that runs through Rome is the Tiber/Nile”) to create an even number of statements for counterbalancing purposes (40 easy, 40 hard). Unlike Experiment 1, fillers were omitted so as to present participants only with either the true/accurate (e.g., “The largest planet in the solar system is Jupiter”) or false/inaccurate (“The largest planet in the solar system is Saturn”) version of each statement. Participants viewed an equal number of true and false statements, and easy and hard statements. They also made equal numbers of interest and accuracy judgments. All participants viewed the same 80 statements, with validity and judgment type counterbalanced across participants. All variations were orthogonal, in that there was the same number of easy/hard, true/false, accuracy judgment/interest judgment items. At test, all participants

³ Given the fully within-subject design of Experiment 2, we recruited 80 participants as an initial examination of effects, which we more thoroughly investigate in Experiment 3.

responded to open-ended questions related to the 80 statements (e.g., “What is the largest planet in the solar system?”), as in Experiment 1.

Procedure

During the exposure phase, participants viewed each of the 80 statements one-at-a-time. A single judgment rating was required for each statement, either for interest (“How interesting was the statement you just read?”) on a scale from 1 (*very uninteresting*) to 6 (*very interesting*), or accuracy (“How accurate was the statement you just read?”) on a scale from 1 (*definitely false*) to 6 (*definitely true*). Statements were presented in a unique random order for each participant, and the order of accuracy and interest judgments was intermixed and randomized. Immediately after, during the test phase, participants received the questionnaire instructions as in Experiment 1, answering 80 open-ended questions related to the earlier read statements. Questions were presented one-at-a-time in a unique random order for each participant.

Results

Questionnaire coding

Responses ($N = 6,240$) were coded as correct, incorrect lure, incorrect other, or unsure/blank. Two raters independently coded a quarter of the responses in the data set, with the remaining coded by one rater only. Inter-rater reliability for dual-coded responses was reliably high ($\kappa = .93$), with all disagreements resolved through discussion.

Models

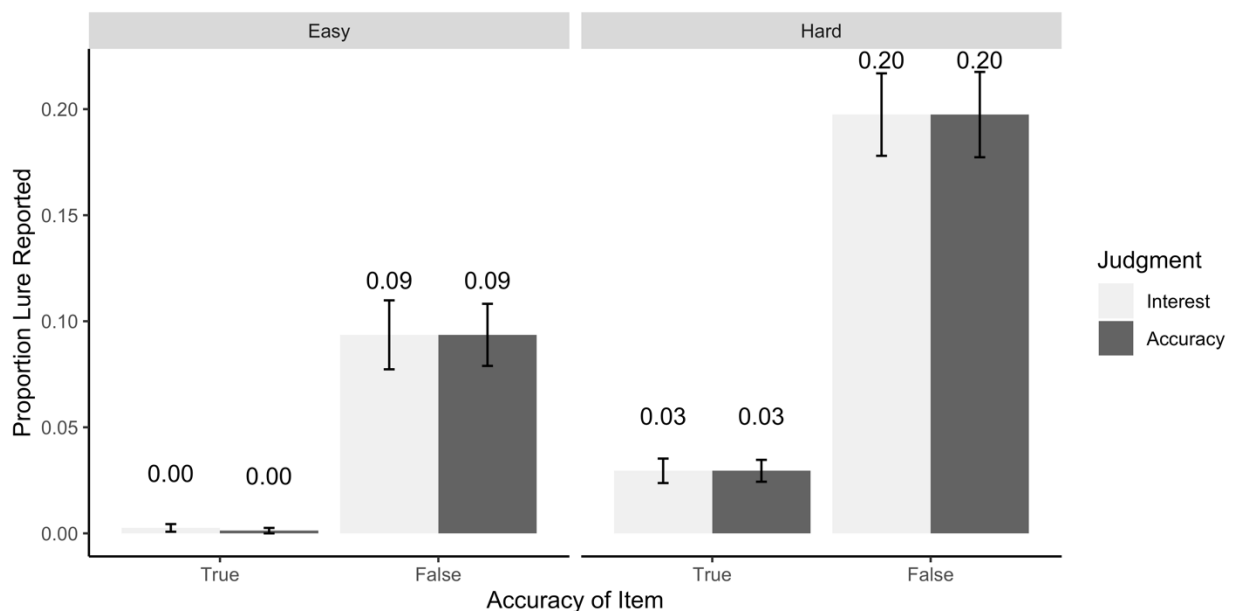
Data were analyzed using the same GLMM model specifications as in Experiment 1, with separate examinations of incorrect lure and correct productions. Model specifications and outputs for Experiment 2 are publicly available at <https://rpubs.com/nsalovich/evaluation-exp2>.

Incorrect Lure Productions

Incorrect lure response rates are summarized in Table 2 and Figure 3. As in Experiment 1, participants again produced more incorrect lures to hard ($M = .11$, $SD = .15$) than to easy questions ($M = .05$, $SD = .11$), $b = 2.04$, $z = 4.97$, $p < .001$. They also reproduced more incorrect lures after reading false statements ($M = .15$, $SD = .16$) than they spontaneously produced those incorrect lures after reading true statements ($M = .02$, $SD = .04$), $b = 2.95$, $z = 9.22$, $p < .001$. There was also a statement difficulty x statement validity interaction, $b = -1.80$, $z = -3.00$, $p = .003$. Simple contrasts revealed that there was a greater difference in incorrect lures produced after reading false versus true information when the statements were hard, $z = 9.91$, $p < .001$, as compared to easy, $z = 6.58$, $p < .001$. There were no main effects or interaction as a function of judgment type, $ps > .05$.

Figure 3

Error rates in Experiment 2



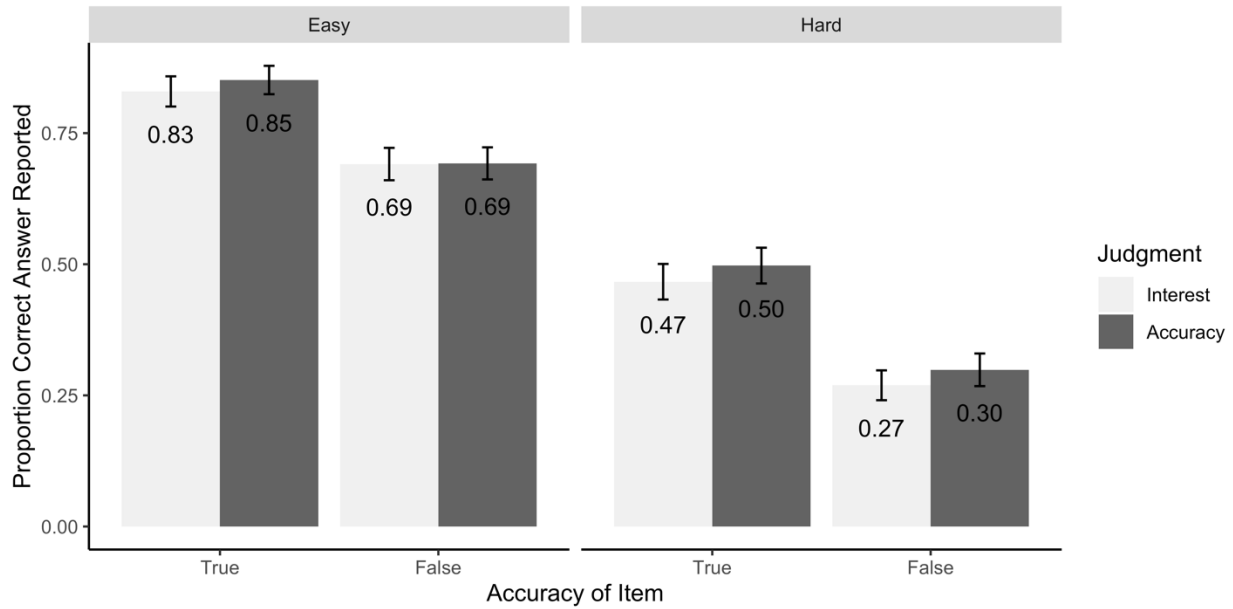
Note: Error rates at test split by validity of statement at exposure (true or false), difficulty of statement at exposure (easy or hard), and judgment type (interest or accuracy). Judgment type (interest or accuracy) was manipulated within subjects. Error bars represent standard error.

Correct Productions

Correct response rates are summarized in Table 2 and Figure 4. As in Experiment 1, participants produced more correct responses for easy ($M = .77$, $SD = .27$) than hard questions ($M = .38$, $SD = .30$), $b = 2.63$, $z = 11.95$, $p < .001$, and more correct responses after reading true ($M = .66$, $SD = .33$) as compared to false statements ($M = .49$, $SD = .34$), $b = 1.40$, $z = 9.34$, $p < .001$. The statement difficulty x statement validity interaction was marginally significant, $b = .29$, $z = 1.80$, $p = .07$, suggesting participants were more likely to produce correct answers after exposure to true than to false statements when the statement contents were easy versus hard. As with the inaccurate reproductions, there were no main effects or interaction as a function of judgment type, $ps > .05$.

Figure 4

Correct response rates in Experiment 2



Note: Correct response rates at test split by validity of statement at exposure (true or false), difficulty of statement at exposure (easy or hard), and judgment type (interest or accuracy). Judgment type (interest or accuracy) was manipulated within subjects. Error bars represent standard error.

Table 2

Mean rates of incorrect lure and correct response production in Experiment 2. Judgment type varied within subjects.

Statement Validity	Statement Difficulty	Lure	Correct
<i>Interest Judgments</i>			
True	Easy	.00 (.02)	.83 (.25)
	Hard	.03 (.05)	.47 (.30)
False	Easy	.09 (.14)	.69 (.27)
	Hard	.20 (.17)	.27 (.25)
<i>Accuracy Judgments</i>			
True	Easy	.00 (.01)	.85 (.24)
	Hard	.03 (.05)	.50 (.30)
False	Easy	.09 (.13)	.69 (.27)

Hard	.20 (.18)	.30 (.28)
------	-----------	-----------

Note: Numbers in parentheses are standard deviations.

Discussion

Participants reproduced more incorrect lures and produced fewer correct answers after exposure to false as compared to true statements, replicating Experiment 1. People were especially likely to produce incorrect lures following exposure to false statements that participants were less likely to already know the answers to, and to produce correct answers following exposure to true information when the statement contents were familiar. We observed no difference in incorrect reproductions or correct responses as a function of the judgment that participants were asked to make while reading the statements. Participants produced similar rates of correct and incorrect lure responses regardless of whether they judged the accuracy or interest of the information at exposure. This lack of difference is consistent with the notion that intermixing judgments led to broad rather than item-specific effects. However, because the experiment did not include control comparisons to an unmixed judgment condition, the data are unclear with respect to whether an evaluative focus associated with accuracy judgments extended to all responses, or a non-evaluative focus associated with interest judgments was broadly influential. That said, the means from Experiment 2 appear more similar to the incorrect lure and correct response rates obtained in the accuracy than interest judgment conditions from Experiment 1. For example, participants in Experiment 2 reproduced incorrect lures from easy, false statements approximately 9% of the time, nearly identical to the proportion of lures reproduced for the same statements in the accuracy judgment condition in Experiment 1 (8%),

and nearly half the rate of the interest judgment condition in Experiment 1 (15%).⁴ This cross-experimental comparison is suggestive but warrants direct test, which we conducted in Experiment 3.

Experiment 3

The purpose of Experiment 3 was two-fold. First, we aimed to replicate the benefits of deliberate evaluation obtained in Experiment 1, predicting that participants who judged statement accuracy at exposure would reproduce fewer incorrect lures and produce more correct responses to questions than would participants who made interest judgments at exposure. Second, we directly examined whether intermixing accuracy and interest judgments would confer overall benefits. We predicted that participants in mixed judgment and accuracy-only conditions would reproduce fewer inaccuracies and produce more correct responses than would participants in an interest-only condition. This would strengthen confidence with respect to the benefits of evaluation on exposures to information, and support mindset claims in the extant literature.

Method

Participants

Two-hundred and forty Amazon Mechanical Turk workers participated in the study, again with the aim of recruiting approximately 80 subjects in each of the three between subject groups. After applying the same exclusion criteria as in the previous experiments, 224 participants (96 female; M age = 38.70 years) remained in the sample ($n = 76$ made all accuracy judgments, $n = 77$ made all interest judgments, $n = 71$ made a mix of accuracy and interest

⁴ Cross-experimental, exploratory GLMMs comparing response rates between the accuracy judgment condition in Experiment 1, interest judgment condition in Experiment 1, and mixed accuracy and interest condition in Experiment 2 affirmed the numerical pattern. Differences emerged between Experiment 2 and interest condition in Experiment 1, but not Experiment 2 and the accuracy condition in Experiment 1.

judgments) totaling 17,920 observations across participants. None participated in Experiments 1 or 2, and all were paid equal to or above the U.S. minimum wage at the time of data collection.

Design

The experiment followed a 2 (statement validity: true or false) x 2 (statement difficulty: easy or hard) x 3 (judgment group: interest-only, accuracy-only, mixed interest and accuracy) mixed design. Statement validity and difficulty were manipulated within subjects, and participants were randomly assigned to one of three judgment groups.

Materials and Procedure

The same materials and procedures were used as in Experiment 1 for the accuracy-only and interest-only conditions, and as in Experiment 2 for the mixed condition. Eighty statements were used and filler items were not included, as in Experiment 2.

Results

Questionnaire coding

All responses ($N = 17,920$) were coded using the same procedure and criteria as in Experiments 1 and 2. Inter-rater reliability for dual-coded responses was reliably high ($\kappa = .95$) with all disagreements resolved through discussion.

Models

All data were again analyzed using GLMMs, and incorrect lure and correct response rates were analyzed separately. As we were specifically interested in whether evaluation reduces the influence of inaccurate content, and given participants rarely produced incorrect lures after viewing true statements as established in Experiment 1, we focused analyses on responses following exposure specifically to false statements. As in Experiment 2, there was no difference in rates of incorrect lure productions in the mixed judgment condition regardless of whether

participants had made an accuracy or interest judgment for the particular statement, $p > .05$ (see Table 3). Therefore, rather than separately examining responses following interest and accuracy judgments in the mixed judgment group, to increase power all responses in the mixed group were jointly compared to both other judgment groups.

In both models, judgment group (accuracy-only, interest-only, and mixed judgments), item difficulty (easy or hard; dummy coded, easy = 1 in both models), and their interaction term were entered as fixed effects. We set the interest-only judgment group as the referent condition for analyses, which allowed us to test whether responses in the accuracy-only and in the mixed judgment groups differed from the interest-only group (e.g., see Weil & Mudrik, 2020 for a similar analysis).⁵ Participants and items were included as random intercepts, and item difficulty and judgment group as random slopes. Model specifications and outputs for Experiment 3 are publicly available at <https://rpubs.com/nsalovich/evaluation-exp3>.

Incorrect Lure Reproductions

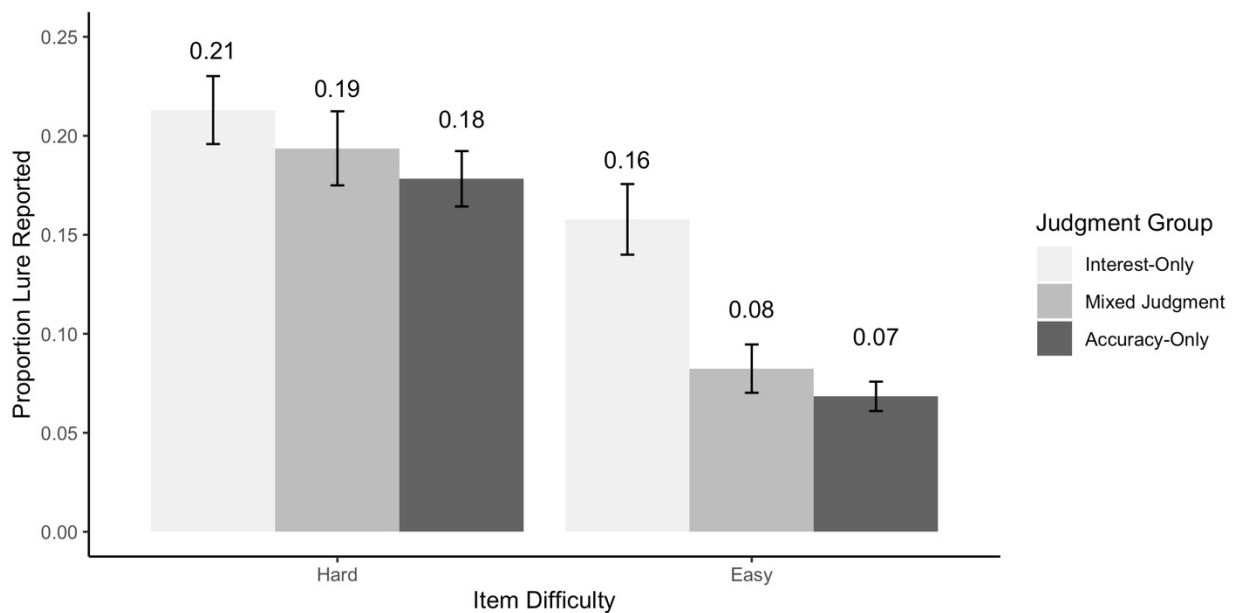
Incorrect lure reproduction rates appear in Table 3. As in the previous experiments there was a main effect of item difficulty, as participants reproduced more incorrect lures for hard ($M = .20$, $SD = .14$) than easy items ($M = .10$, $SD = .12$), $b = -.52$, $z = -2.05$, $p = .04$. As predicted, a significant statement difficulty x judgment group interaction emerged when comparing accuracy-only and interest-only groups, $b = -.76$, $z = -3.41$, $p < .001$, as well as mixed judgment and interest-only groups, $b = -.75$, $z = -3.32$, $p < .001$. Simple contrasts revealed that participants reproduced fewer incorrect lures when accuracy judgments were required, even when intermixed with interest judgments, as compared to in the interest-only condition, specifically when

⁵ As recommended by a reviewer, we also ran exploratory analyses for both incorrect lure and correct answer responses using *post hoc* orthogonal contrasts, and found effects consistent with the results and interpretation reported here. The specifications and outputs of these analyses can be found along with the rest of the model specifications and outputs for Experiment 3 at <https://rpubs.com/nsalovich/evaluation-exp3>.

statements were easy (accuracy only: $z = -4.61, p < .001$; mixed judgment: $z = -4.24, p = .001$), but not hard (accuracy only: $z = -1.31, p = .39$; mixed judgment: $z = -.99, p = .58$). There was no difference in incorrect lure reproductions regardless of whether participants always made accuracy judgments or made them only half of the time (see Figure 5).

Figure 5

Error Rates in Experiment 3



Note: Error rates following false information split by statement difficulty (easy or hard) and judgment group (interest-only, mixed judgment, or accuracy-only). In the mixed judgment group, interest and accuracy judgments were made within subjects and reduced into a single bar. Error bars represent standard error.

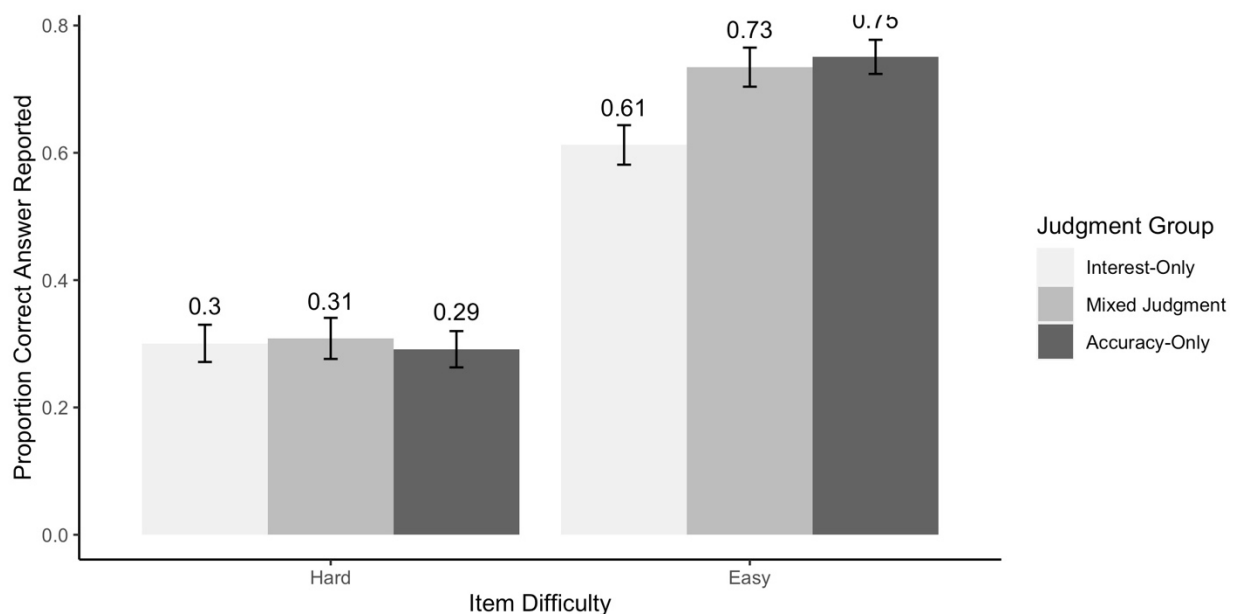
Correct Productions

Correct response rates are summarized in Table 3. Participants produced more correct responses to easy ($M = .70, SD = .26$) than hard items ($M = .30, SD = .26$), $b = 2.01, z = 8.67, p <$

.001. Replicating Experiment 1, a statement difficulty \times judgment group interaction emerged when comparing the accuracy-only and interest-only groups, $b = .87$, $z = 3.32$, $p < .001$. Simple contrasts revealed that accuracy judgments encouraged correct answers when the statements were easy, $z = 3.23$, $p = .004$, but not hard, $z = -.06$, $p = .99$. A statement difficulty \times judgment group interaction also emerged comparing correct productions from the mixed judgment to the interest-only judgment group, $b = .74$, $z = 2.75$, $p = .006$, with analysis of simple contrasts again demonstrating benefits for easy statements, $z = 3.01$, $p = .007$, but not hard statements, $z = .26$, $p = .96$. In sum, participants instructed to make only accuracy judgments, or accuracy judgments intermixed with interest judgments, produced more correct answers as compared to participants instructed to only make interest judgments. The interaction effects indicate this benefit emerged for easy but not hard statements (see Figure 6).

Figure 6

Correct Response Rates in Experiment 3



Note: Correct response rates following false information split by statement difficulty (easy or hard) and judgment condition (interest-only, mixed judgment, or accuracy-only). In the mixed judgment group, interest and accuracy judgments were made within subjects and reduced into a single bar. Error bars represent standard error.

Table 3

Mean rates of incorrect lure and correct response production in Experiment 3. Judgment type varied between and within subjects depending on condition.

Statement Validity	Statement Difficulty	Lure	Correct
Mixed Judgment (Within Subjects)			
<i>Interest Judgments</i>			
True	Easy	.01 (.03)	.86 (.25)
	Hard	.04 (.06)	.47 (.32)
False	Easy	.09 (.14)	.72 (.27)
	Hard	.19 (.18)	.29 (.29)
<i>Accuracy Judgments</i>			
True	Easy	.00 (.02)	.89 (.22)
	Hard	.03 (.05)	.53 (.34)
False	Easy	.07 (.11)	.75 (.28)
	Hard	.19 (.17)	.32 (.29)
Separate Judgment (Between Subjects)			
<i>Interest Judgments</i>			
True	Easy	.00 (.01)	.82 (.23)
	Hard	.02 (.03)	.53 (.26)
False	Easy	.16 (.16)	.61 (.27)
	Hard	.21 (.15)	.30 (.26)
<i>Accuracy Judgments</i>			
True	Easy	.01 (.02)	.85 (.24)
	Hard	.03 (.04)	.52 (.29)
False	Easy	.07 (.06)	.75 (.23)
	Hard	.18 (.12)	.29 (.25)

Note: Numbers in parentheses are standard deviations.

Discussion

Experiment 3 replicated the results of Experiment 1: Participants who judged statement accuracy at exposure reproduced significantly fewer incorrect lures at test as compared to participants who judged statement interest at exposure. Participants in the accuracy-only judgment condition also produced more *correct* responses following exposure to false information than did participants in the interest-only condition. The decrease in incorrect lure reproductions and increase in correct productions was only observed for easy statements that were likely familiar to participants. This is indicative of accuracy judgments encouraging the availability and use of correct prior knowledge over recently read inaccuracies. Given the hard statements were likely unfamiliar, participants would be unable to leverage prior knowledge to recognize them as false. Thus, evaluative benefits did not emerge for hard, false information. Experiment 3 additionally demonstrated that the benefits obtained when participants were asked to consider accuracy throughout the task were also observed when participants were asked to make accuracy judgments for a random half of the statements. As shown in Figures 5 and 6, the rates of incorrect lure reproductions and correct answers produced in the mixed judgment and accuracy-only conditions were nearly identical. Being asked to consider accuracy, even intermittently, encouraged an evaluative mindset for the task generally, rather than only to items that specifically prompted such contemplation.

General Discussion

This chapter includes three studies investigating the role of deliberate evaluation in reducing the problematic consequences associated with exposures to false information. Across three experiments, participants who were instructed to rate the accuracy of information

reproduced fewer inaccurate ideas on post-reading questions than did participants who rated their interest in the information. These results extend prior work on the benefits of “fact-checking” (Brashier et al., 2020; Rapp, Hinze, et al., 2014), establishing that an accuracy focus can reduce people’s actual use of inaccuracies to answer open-response questions. Accuracy judgments also increased the likelihood that participants would provide correct answers to questions by relying on accurate existing knowledge rather than on the false information they read. Any skepticism potentially instantiated by the evaluative instructions with respect to inaccurate information did not extend to their considerations of accurate knowledge. Other tasks using instructions to encourage skepticism have at times problematically resulted in participants becoming overall more conservative in their responses (Andrews-Todd et al., 2021; Marsh & Fazio, 2006; Salovich et al., 2021). In the current project, such responses could have resulted in participants leaving easy questions they presumably already knew the answer to blank. Evaluating statements for accuracy instead beneficially increased the likelihood that participants would leverage their accurate prior knowledge when available, providing correct answers in lieu of reproducing inaccuracies.

These findings are consistent with the idea that evaluative benefits emerge by reducing the extent to which presented inaccuracies interfere with people’s correct understandings. During exposure to false information, concepts may still be activated and integrated during comprehension given the routine operations of memory (Anderson, 1981; Cook & O’Brien, 2014; O’Brien & Cook, 2016b; Lewis & Anderson, 1976). However, instructions to engage in deliberate evaluation can disrupt these moment-by-moment processes as accurate prior knowledge becomes simultaneously activated during exposures to inaccurate information. According to memory-for-change (Negley et al., 2018; Jacoby & Wahlheim, 2013; Wahlheim,

2015; Wahlheim & Jacoby, 2013) and RI-Val (Cook & O'Brien, 2014; O'Brien & Cook, 2016b) accounts, detecting inaccuracies requires retrieving and activating correct prior knowledge alongside presented falsehoods. For example, to effectively evaluate the statement "Narcolepsy is the term for an inability to sleep," people must leverage their existing knowledge (e.g., retrieve the fact that this definition actually refers to "insomnia") to determine whether the claim is true or false. As a result, the correct answer can also be included into any encoded representation of the stated claim, such that the accurate prior knowledge explicitly competes with the presented inaccuracy. This, in turn, should reduce reliance on the inaccuracy in favor of the correct idea. In line with the effects of retrieval practice, where the act of recalling information from memory supports its retrieval on later tasks (Bjork, 1988; Roediger & Butler, 2011), evaluation can encourage in-the-moment recall to enhance the subsequent availability of correct answers. Evaluation therefore supports encoding and retrieval processes, along with affording critical engagement with content.

This mechanistic explanation aligns with other models describing the processes and conditions underlying effective knowledge updating and revision. According to the Knowledge Revision Components (KReC) Framework (Kendeou & O'Brien, 2014), overcoming false ideas and updating existing worldviews requires the co-activation of accurate ideas with outdated or incorrect understandings. These accounts highlight that while prior understandings are unlikely to be completely eliminated, co-activation can increase the likelihood that correct information is privileged over inaccurate, outdated memory traces (e.g., Braasch & Bråten, 2017; O'Brien & Cook, 2016b; van den Broek et al., 2005). Co-activation therefore proves crucial for explaining and enacting the benefits that evaluations offer for combatting inaccurate exposures.

These results offer fodder for considering how detection and correction of inaccuracies can result in disparate downstream effects. A number of studies have documented that people are sensitive to information that contradicts previous presentations and prior knowledge (Albrecht & O'Brien, 1993; Cook & O'Brien, 2014; Richter et al., 2009; Singer, 2006, 2013). For example, readers slow down when they encounter false claims and ideas in a text (e.g., that George Washington was not elected the first president of the United States; Rapp, 2008). Yet studies describing people's detection of erroneous information should not necessarily be equated with people's in-the-moment corrections of that information. For example, the degree to which prior knowledge needs to be activated to determine whether a given claim is false may depend on the complexity of the claim. Consider that Weil and Mudrik (2020) reported instances in which people evaluated statements by detecting semantic mismatches within sentence content seemingly without reliance on correct prior knowledge. Detecting such obvious falsehoods (e.g., "A fire is cold"; "Blood is green"; "Cars have legs") may invoke processes associated with the detection of semantic anomalies (e.g., responding to a lack of semantic associates between "fire" and "cold"; see Kutas & Hillyard, 1980; Sanford et al., 2011); more subtle or complicated inaccuracies referencing the names of famous people in history or the location of capital cities (as used in statements from the current experiments) may require more effortful evaluation and at times even strategic searches of memory to recognize that they are false. As such, people may be able to detect some inaccuracies without adopting an explicit goal or focus to evaluate (Isberner & Richter, 2014a), yet still rely on them if detection fails to activate correct prior knowledge (e.g., Weil & Mudrik, 2020), or due to lingering uncertainty about what is true (Rapp & Salovich, 2018; Salovich et al., 2021). Detection on its own is thus likely insufficient for disrupting the downstream effects of inconsistencies or inaccuracies (e.g., Marsh & Fazio, 2006;

Rapp, Hinze, et al., 2014). Careful examination of both nonstrategic and goal-driven processes may help elucidate when and why these benefits appear specific to deliberate evaluations.

The current work additionally provides insight into what is needed, and when, to establish and maintain an evaluative mindset. Evaluations and evaluative mindsets have been of specific interest in work intended to encourage critical thinking during experiences with inaccurate information (e.g., Bago et al., 2020; Pennycook & Rand, 2019; Mayo, 2015; Schul & Mayo, 2014). While people can certainly engage in evaluation when provided explicit and/or consistent prompting and reminders (e.g., Brashier et al., 2020; Rapp, Hinze, et al., 2014), a thornier issue involves determining conditions under which people consider the accuracy of information *without being asked to do so* (Isberner & Richter, 2014b; Mayo, 2019; Wiswede et al., 2013). Relevant discussions have focused on the likelihood and role of automatic evaluations, with some evidence suggesting that people routinely validate information as part of comprehension, and other evidence asserting that evaluation occurs only under specific circumstances (Isberner & Richter, 2014b, Richter, 2015; Richter, et al., 2009; Singer, 2006, 2013). In the current study, we observed benefits when participants were explicitly instructed to evaluate all of the statements they were asked to read (Experiment 1), and when half of the statements in a larger set required evaluation (Experiment 2 and 3). Notably, the evaluations prompted by the task were beneficial for other items that required non-evaluative judgments. These data cannot speak to whether participants strategically and purposively enacted evaluative judgments even when tasked with interest considerations, or if they confabulated accuracy and interest judgments in ways that benefitted performance. Rather, the results suggest that instructions to evaluate can assert an influence beyond the circumstances under which they are required, as associated with evaluative mindsets. We acknowledge such a context is a far cry

from spontaneous evaluation devoid of accuracy prompts; but the observation that participants became overall *more* rather than *less* accurate in their responses when evaluative and non-evaluative judgments were intermixed is noteworthy for establishing understandings of when and how to promote routine considerations of accuracy, and more generally for mindset accounts.

Future research should continue to examine the kinds of prompts that encourage and maintain evaluative mindsets. Recent work has argued that brief accuracy “nudges,” in which people are asked to consider the accuracy of a single or small number of ideas prior to exposure to false information, can decrease people’s sharing intentions for that information (Pennycook, Epstein, et al., 2021; Pennycook et al., 2020). The current study differed from those projects in two ways: First, the task involved a protracted activity in which participants were instructed to reflect on accuracy throughout the task rather than only once prior to it, with the intermixed conditions demonstrating analogous benefits as reported in “nudge” work. Second, the task examined people’s reproductions of text content rather than sharing inclinations, demonstrating that evaluative mindsets are useful for conditions involving actual knowledge application. These latter benefits might have been due specifically to the recurring requirement to consider evaluations across the task, as a recent pre-registered study failed to replicate “nudge” benefits for sharing intentions and as observed with the passage of time (Roozenbeek et al., 2021). People may require explicit and/or frequent reminders to obtain more substantial, long-lasting benefits than are afforded by brief attempts at intervention (Salovich & Rapp, 2021; Sparks & Rapp, 2011). Follow-up studies could further investigate the transfer and durability of such effects, perhaps by using alternating blocks of evaluative and non-evaluative judgments as leveraged in interleaving and spacing studies (e.g., Chen et al., 2021), or by introducing time delays between exposures and test. Brashier et al. (2020) recently reported that the benefits of an accuracy focus

persisted two days after initial exposure to the statements, suggesting potential longer-term consequences of evaluations for reinforcing correct understandings.

In Chapter 2, we continue to explore the conditions that lead to evaluative mindsets, testing whether people can be encouraged to adopt an accuracy focus during comprehension without direct instruction to do so. Specifically, we expand on previous work wherein people's considerations of their own susceptibility to false information have been shown to play a crucial role in enacting processes and strategies related to evaluation (Salovich & Rapp, 2021). In Experiment 4, we consider whether and how positive or negative performance feedback about people's own susceptibility to inaccurate information affects their uptake and use of presented inaccuracies. We build on this idea in Experiment 5, where we examine whether an awareness that one's fallibility to inaccurate information is being monitored can also motivate evaluative mindsets.

Chapter 2:

Using Feedback and Monitoring to Reduce the Influence of False Information

Exposure to false information can problematically influence people's understandings, even when it blatantly contradicts people's accurate and accepted prior knowledge (e.g., Fazio, Brashier, et al., 2015; Fazio et al., 2019; Rapp, 2008; Rapp, 2016). This well-replicated finding connects to concerns about the prevalence and consequences of inaccurate information conveyed through social media, gossip, and other disintermediated channels (Britt et al., 2019; Lewandowsky et al., 2017). Recent work shows that explicitly asking participants to "fact-check" the accuracy of information during reading reduces their likelihood of reproducing and sharing false information, as compared to participants who read the same content without evaluative prompting (Brashier et al., 2020; Pennycook, Epstein, et al., 2021; Rapp, Hinze, et al., 2014). Salovich, Kirsch, et al. (2022) recently exemplified this benefit, tasking participants with rating true and false statements (e.g., "Jupiter/Saturn is the largest planet in the solar system") for accuracy (i.e., an evaluative judgment) or interest (i.e., a non-evaluative judgment). On post-reading questions, participants who evaluated statements for accuracy were less likely to reproduce ideas from the false statements ("Saturn"), and more likely to rely on correct prior knowledge than were participants who made interest judgments. People may fail to evaluate the accuracy of information during comprehension, leading to competition and/or interference between recently encoded false information and accurate prior knowledge. Explicit evaluation during reading disrupts these interference effects by encouraging the activation of relevant prior knowledge. This increases the likelihood of detecting false information when it contradicts what is already known to be true, and increases the availability of accurate prior knowledge for use on subsequent tasks.

Despite replications of these benefits, it remains an open question whether and how people can be encouraged to adopt such practices. Researchers have begun investigating the possibility of prompting mindsets that might sustain evaluative processing even during non-evaluative tasks (Mayo, 2015, 2019). This would be particularly beneficial as non-evaluative tasks are associated with shallow processing and comprehension, and fail to encourage activation of relevant prior knowledge (Hawkins & Hoch, 1992; Hawkins et al., 2001), supporting the uptake of inaccurate ideas. Thankfully, research suggests people are amenable to adopting an evaluative focus during reading; for example, participants who were intermittently prompted to consider the accuracy of information showed reductions in their reproductions of blatant falsehoods to answer post-reading questions, and at levels commensurate with consistent prompting (Salovich, Kirsch, et al., 2022). Other projects have used accuracy nudges to encourage skepticism, resulting in benefits including reductions in sharing fake news headlines (Pennycook, Epstein, et al., 2021; Pennycook et al., 2020, but see Roozenbeek et al., 2021).

These projects have tended to rely on explicit instructions to prompt evaluation, which may not be easy to enact in contemporary information environments, or practiced by people given their diverse reasons for engaging with information. Recent work has examined how people's thoughts and beliefs about their susceptibility to the influence of inaccurate information might also be leveraged to encourage evaluation (e.g., Rapp & Salovich, 2018; Salovich et al., 2021). People often seem unaware of this susceptibility, overestimating their ability to discern truth from falsehood (Lyons et al., 2021), and underestimating the importance of evaluation for learning and comprehension (Salovich & Rapp, 2021). Making people aware of their susceptibility can be beneficial: Participants asked to recollect times they mistakenly relied on false ideas were less likely to be influenced by presented inaccuracies on a subsequent task than

were participants who did not engage in such recollection (Salovich & Rapp, 2021). These results suggest that drawing attention to a propensity towards “falling for” false information might encourage attention to and use of evaluative comprehension strategies.

Several mechanisms may underlie the benefits associated with drawing attention to this susceptibility. One possibility is that people may fail to recognize the difficulties they can personally experience from being exposed to falsehoods, so confronting them with examples of their problematic past reliance might encourage them to consider information carefully. Another possibility is that prompts to consider potential consequences of such exposures could motivate people to process information more critically than they might otherwise. Determining whether one or both of these possibilities is viable proves informative for identifying when negative effects resulting from inaccurate exposures might emerge, and for designing interventions intended to encourage evaluation.

The current project sought to disentangle these possibilities. In two experiments we examined whether and how confronting people with their susceptibility to false information might reduce its influence. Participants made non-evaluative interest ratings about true and false statements and then answered related questions. In Experiment 4, we tested whether providing positive or negative performance feedback after answering these questions would influence participants’ use of newly presented inaccuracies on a new set of questions, as compared to when feedback was not provided. In Experiment 2, we examined whether making participants aware that their susceptibility to false information was being monitored would influence their use of inaccuracies to answer questions. The findings highlight the critical role of people’s considerations about their resistance to inaccuracies in accounts of the influence of inaccurate exposures, and suggest methods for encouraging evaluation in practical settings.

Experiment 4

Experiment 4 examined whether providing feedback concerning people's susceptibility to false information would affect their use of newly presented inaccuracies to answer related questions. Participants read true (e.g., "An inability to sleep is called insomnia") and false statements (e.g., "An inability to sleep is called narcolepsy"), half of which were easy and likely well-known to participants (e.g., "Young sheep are called lambs/calves"), and half hard and less likely to be known (e.g., "The river that runs through Rome is the Tiber/Nile"). Participants were instructed to rate how interesting each statement was to encourage attention to and comprehension of the information without necessitating considerations of accuracy (see Brashier, et al., 2020; Calvillo & Smelter, 2021; Salovich, Kirsch, et al., 2022 for similar procedures). Next, participants answered questions related to the statements (e.g., "What is the medical term for the inability to sleep?"). Afterwards they were randomly assigned to receive feedback indicating how their responses were influenced by earlier presented false information. Unlike feedback given for specific items (i.e., informing participants that a particular answer is correct or incorrect; Marsh et al., 2016; Pashler et al., 2005), summative, relative feedback was delivered after participants completed the activity. One-third of participants received negative feedback, being told they were influenced more than the average person; one-third received positive feedback, being told that they were influenced less than average; and one-third did not receive any feedback. Next, participants read a *new* set of true and false statements and answered questions related to them. This allowed us to measure participants' inaccurate reproductions (referred to as *incorrect lures*; e.g., "narcolepsy") and correct responses (e.g., "insomnia") before and after receiving feedback.

We consider three possibilities for how feedback could affect performance. One hypothesis is that positive and negative feedback might differentially influence people's subsequent responses. Negative feedback indicating susceptibility to inaccurate ideas could encourage evaluation, such that after receiving feedback, participants might provide fewer incorrect reproductions and more correct answers despite exposure to false information. In contrast, positive feedback indicating a healthy resistance to inaccurate ideas might encourage people to "let their guard down" (Hattie & Timperley, 2007; Podsakoff & Farh, 1989), such that they produce more incorrect reproductions and fewer correct answers. This possibility aligns with projects showing associations between overconfidence and a decreased propensity to detect and resist inaccuracies (e.g., Lyons et al., 2021; Salovich & Rapp, 2021).

Another hypothesis is that feedback might be beneficial but not differentially so. Feedback is useful not only for correcting mistakes, but also for increasing effort, motivation, and engagement (Hattie & Timperley, 2007; Sadler, 1989). It can encourage people to reflect on and assess their processing strategies, and to adjust those strategies as necessary depending on task goals (Lee et al., 2010; Roll et al., 2011). Positive and negative feedback both draw attention to the influence of false ideas, which people may not have considered or acted upon during the task up to that point. While negative feedback directly signals the need for evaluative considerations, positive feedback can motivate behavioral changes when a goal state is highly desirable, as may be the case with resisting false information (Altay et al., 2020; Van-Dijk & Kluger, 2000, 2001; Waruwu et al., 2020). Both forms of feedback could encourage evaluation. Therefore, both negative and positive feedback, in contrast to no feedback, might lead to fewer inaccurate reproductions and more correct responses.

A third hypothesis is that feedback might not encourage evaluative considerations given the explicit task of making interest rather than evaluative judgments. But we suspected this was unlikely given participants are amenable to overarching evaluative considerations even when a task randomly requires interest *or* accuracy judgments for different items within a single response activity (Salovich, Kirsch, et al., 2022).

Method

Participants

Two-hundred and forty undergraduates participated in the study in exchange for course credit. In line with previous studies using these materials (e.g., Salovich, Kirsch, et al., 2022), we recruited 80 participants for each of three between subject conditions. This allowed for sufficient observations of each item type both within and across participants to achieve adequate power while reducing Type 1 Error rates for the mixed effect analyses (Luke, 2017). We excluded eight participants who failed a multiple-choice comprehension check or English language check, or reported looking up answers, leaving a sample of 232 participants (118 female; *M* age = 18.71 years), and 18,560 observations across participants.

Design

The experiment followed a 2 (statement validity: true or false) x 2 (statement difficulty: easy or hard) x 3 (feedback type: positive, negative, no feedback) mixed design. Statement validity and difficulty were manipulated within subjects and feedback type was manipulated between subjects.

Materials

We used 80 facts (Salovich, Kirsch, et al., 2022) drawn from previously published general knowledge norms (Tauber et al., 2013). Forty test items were easy (known by most

people) and 40 were hard (known by few people), based on the norms. Each normed fact included an accurate answer and an incorrect lure, and therefore could appear in a true (e.g., “An inability to sleep is called insomnia”) or false version (“An inability to sleep is called narcolepsy”), matched on word length. The 80 declarative statements were evenly split into two sets of 40 statements, each containing 20 easy and 20 hard items. Half of the participants received Set 1 pre-feedback and Set 2 post-feedback, and half received Set 2 pre-feedback and Set 1 post-feedback. Participants responded to open-ended questions related to the 80 statements (e.g., “What is the largest planet in the solar system?”), 40 pre-feedback and 40 post-feedback, from the relevant assigned sets. The full set of materials can be found on OSF (osf.io/cq673/).

Procedure

The experiment involved three phases (for visual depiction of the procedure, see Figure 7). After providing informed consent, participants completed the *pre-feedback phase* in which they rated the first set of 40 true and false general knowledge statements for how interesting they were, on a scale from 1 (very uninteresting) to 6 (very interesting). Statements were presented one-at-a-time in a unique random order for each participant, and consisted of an equal number of easy and hard statements. After rating all statements, participants were asked to complete a general knowledge questionnaire and to “answer based on what you believe to be true about the world.” They answered 40 open-ended questions related to the earlier statements, presented in a unique random order, by typing their answers in a textbox. They were instructed not to look up any answers, and were told that they could leave the box blank or write unsure/no answer if they did not know the answer.

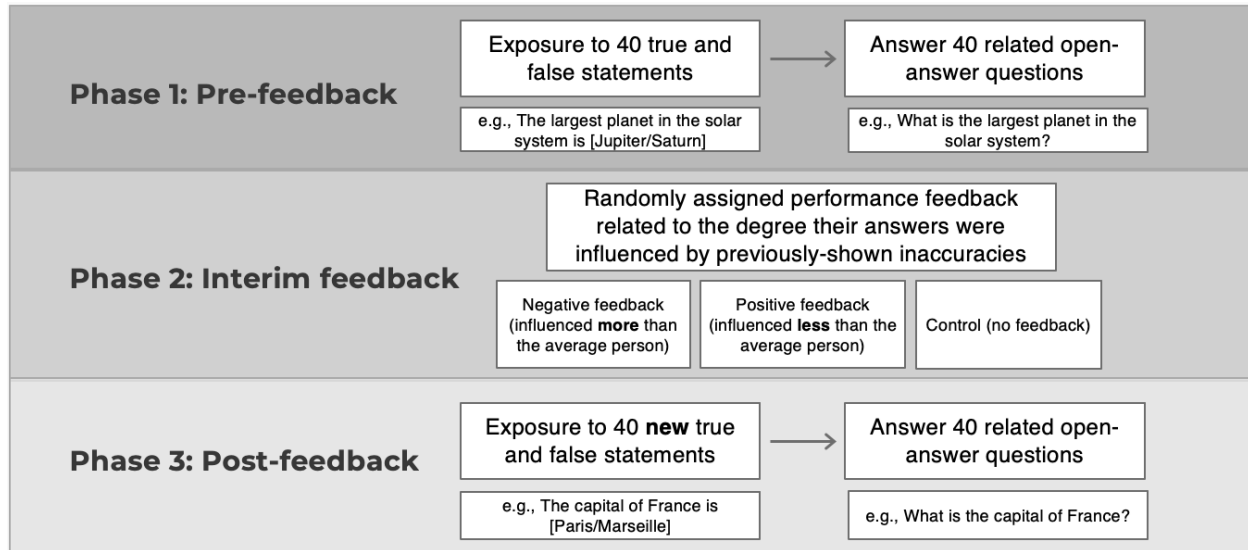
In the *feedback phase*, participants were randomly assigned feedback about how their questionnaire responses were influenced by the earlier presented statements. Participants were

randomly assigned to one of three conditions, such that the feedback they received was not reflective of their actual performance. One-third received negative feedback, being told they were influenced more than the average person completing the task in the past; one-third received positive feedback, being told they were influenced less than the average person; and one-third did not receive any feedback. The full text of feedback is available in Appendix A. We chose to provide participants with feedback regarding their susceptibility to inaccuracies relative to other individuals, as previous work has discussed such comparisons as effective for encouraging behavioral change (e.g., Andrews-Todd et al., 2021; French et al., 2011; Salovich & Rapp, 2021).

The final, *post-feedback phase* tested the effects of the feedback. Participants read and answered questions related to a new set of 40 true and false statements. The procedure was identical to the pre-feedback phase using these new statements and questions. The sets of statements used in the pre- and post-feedback phases were balanced across participants.

Figure 7

Visual Depiction of Experiment 4 Procedure



Results

Questionnaire coding

All responses ($N = 18,560$) were coded as correct (e.g., “insomnia), incorrect lure referring to the presented inaccuracy (e.g., “narcolepsy”), incorrect other (e.g., responding with anything else such as “jet lag” or “sleepwalking”), or unsure/blank. Two raters independently coded a quarter of the responses in the data set, and inter-rater reliability was reliably high ($\kappa = .97$) with all disagreements resolved through discussion. The remaining responses were coded by one rater only. While this paper focuses on rates of incorrect lures and correct responses, all coded data are publicly available on OSF (osf.io/cq673/).

Models

Data analyses were run using generalized linear mixed effect models (GLMM) with the R packages `lme4` (Bates et al., 2015) and `lmerTest` (Kuznetsova et al., 2017). These mixed effect models simultaneously account for the variance introduced due to overall differences across

participants and items (random intercepts), as well as the possibility that manipulations of statement difficulty and accuracy could affect each participant differently (random slopes).

Model specifications and outputs for Experiment 4 are publicly available at <https://rpubs.com/nsalovich/feedback-exp1>.

We also conducted separate analyses of incorrect lure and correct answer responses to test the hypothesized effects, which is consistent with previous research examining how presentations of information affect people's responses to open-answer questions (e.g., Donovan & Rapp, 2020; Kelley & Lindsay, 1993; Fazio et al., 2013; Fazio & Marsh, 2008; Marsh & Fazio, 2006; Marsh et al., 2003; Salovich, Kirsch, et al., 2022). Tables summarizing the distribution of response types for all three experiments are available as supplemental material on OSF (osf.io/cq673/).

Incorrect Lure Productions

Incorrect lure reproduction rates are summarized in Table 1 and Figure 8. To analyze the reproductions, incorrect lure responses were coded as 1 and all other responses as 0. To replicate previous effects, we first examined how statement validity and difficulty affected incorrect lure reproduction. The GLMM included statement validity (contrast coded; true = -.5, false = .5) and statement difficulty (contrast coded; easy = -.5 and hard = .5) as fixed effects, participant and item as random intercepts, and statement validity and difficulty as random slopes. This model simultaneously accounted for variance due to random selection of participants and items, as well as variance in how statement validity and difficulty affected each participant (Richter, 2006).

We observed a main effect of difficulty, with participants reproducing incorrect lures more often to questions about hard ($M = .20$, $SD = .23$) than easy statements ($M = .13$, $SD = .23$), $b = 1.52$, $z = 5.56$, $p < .001$. Participants also reproduced more incorrect lures after reading false

statements containing those inaccuracies ($M = .31$, $SD = .26$) as compared to spontaneously producing incorrect lures after reading true statements ($M = .02$, $SD = .06$), $b = 4.25$, $z = 34.03$, $p < .001$. We also observed a significant statement validity x statement difficulty interaction, $b = -1.53$, $z = -6.25$, $p < .001$. To investigate this interaction, simple contrasts were calculated using the emmeans R package (Lenth, 2019). While incorrect lure reproductions were more likely after exposure to false as compared to true statements, this pattern was more readily observed for hard, $z = 32.82$, $p < .001$, than easy questions, $z = 22.45$, $p < .001$. In other words, people were more influenced by inaccuracies about lesser-known than well-known information. These results replicate previous studies (e.g., Donovan & Rapp, 2020; Hinze et al., 2014; Marsh et al., 2003; Salovich, Kirsch, et al., 2022).

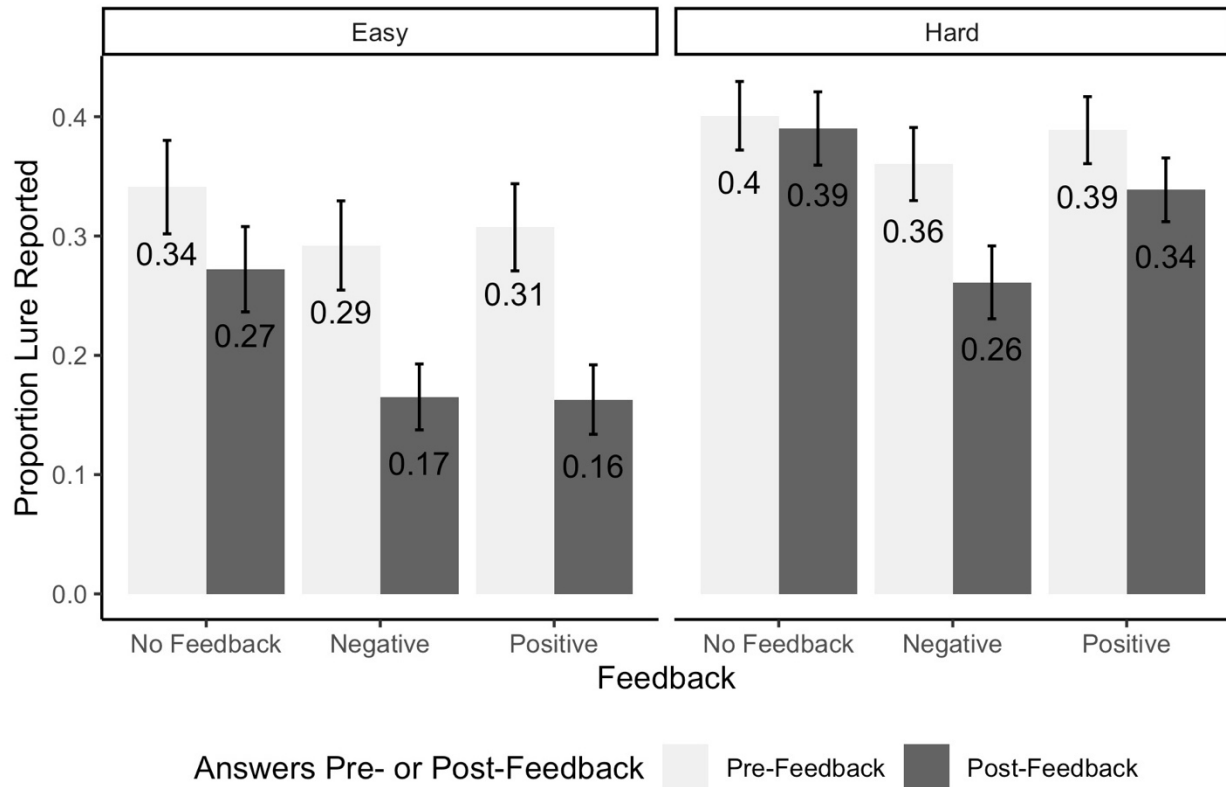
We next assessed how interim feedback affected inaccurate reproductions. We ran a GLMM predicting incorrect lure reproduction following exposure specifically to false statements. Fixed effects were statement difficulty (contrast coded; easy = $-.5$ and hard = $.5$), experiment phase (contrast coded; pre-feedback = $.5$ and post-feedback = $-.5$), and feedback type (no feedback set as the referent condition). Participants and items were again included as random intercepts, and statement difficulty as a random slope.

Participants overall reproduced fewer incorrect lures in the post-feedback phase ($M = .26$, $SD = .25$) as compared to the pre-feedback phase ($M = .35$, $SD = .27$), $b = .32$, $z = 3.38$, $p = .007$. There was also significant experiment phase x statement difficulty interaction, $b = -.43$, $z = -2.29$, $p = .02$, with simple contrasts revealing the reduction in lure reproductions between phases occurred more for easy, $z = 10.78$, $p < .001$, than hard statements, $z = 4.43$, $p < .001$. Across both study phases, participants who received negative interim feedback ($M = .27$, $SD = .26$) reproduced fewer incorrect lures than did participants in the no feedback group ($M = .35$, $SD =$

.27), $b = -.60$, $z = -2.52$, $p = .01$, with no difference in lure reproduction between positive ($M = .30$, $SD = .25$) and control feedback groups, $b = -.34$, $z = -1.53$, $p = .13$. However, critically, significant experiment phase x feedback type interactions indicated that the decrease in lure reproduction from pre- to post-feedback was greater for participants who received either negative feedback, $b = .52$, $z = 3.79$, $p < .001$ or positive feedback, $b = .39$, $z = 2.86$, $p = .004$, as compared to participants who received no interim feedback. This was confirmed by follow-up analyses of simple contrasts. The negative and positive feedback group differed from the no feedback group with respect to lure reproductions in the post-feedback experimental phase, (*positive feedback-no feedback*: $z = 3.44$, $p = .002$; *negative feedback-no feedback*: $z = 2.28$, $p = .06$), but did not differ in the pre-feedback phase, $ps > .05$.

Although neither experiment phase x feedback type x difficulty interaction reached significance, $ps > .05$, feedback qualitatively reduced incorrect lure reproductions more so for easy as compared to hard items (see Figure 2). To interrogate this pattern of effects, we ran a pairwise contrast to compare the experiment phase x feedback type interaction between easy and hard items. As expected, both positive feedback, $z = 2.70$, $p = .007$, and negative feedback groups, $z = 2.89$, $p = .004$, showed reductions in lure reproductions relative to the no feedback group when statements were easy. However, for hard statements, differences only emerged between the negative feedback group and the no feedback group, $z = 2.46$, $p = .02$. As compared to when no feedback was received, positive feedback reduced incorrect lure reproduction only when information was familiar, but negative feedback reduced lure reproduction for both familiar and unfamiliar information (see Figure 8).

Figure 8

Error Rates Following False Information in Experiment 4

Note: Error rates split by difficulty of statement at exposure (easy or hard) and feedback type (no feedback, negative feedback, or positive feedback). Light gray bars represent responses made prior to receiving interim feedback and dark gray bars represent responses made after receiving interim feedback. Error bars represent standard error.

Correct Answer Productions

Correct response rates are summarized in Table 1 and Figure 9. For this analysis, responses that were correct were coded as 1 and all other responses as 0. We first examined whether statement validity and difficulty affected correct responses. The GLMM included statement validity (contrast coded; true = .5 and false = -.5) and difficulty (contrast coded; easy =

.5 and hard = -.5) as fixed effects, participant and item as random intercepts, and statement validity and difficulty as random slopes.

Participants were more likely to produce correct responses to questions about easy ($M = .77$, $SD = .31$) than hard statements ($M = .32$, $SD = .28$), $b = 3.14$, $z = 15.06$, $p < .001$. They also produced more correct responses after exposure to true ($M = .70$, $SD = .31$) as compared to false statements ($M = .39$, $SD = .35$), $b = 2.45$, $z = 46.51$, $p < .001$. There was also a significant statement validity x statement difficulty interaction, $b = .97$, $z = 9.28$, $p < .001$. Simple contrasts revealed that participants were more likely to produce correct answers after exposure to true as compared to false statements when the statement contents were easy, $z = 34.02$, $p < .001$, versus hard, $z = 32.62$, $p < .001$.

We next assessed whether interim feedback affected correct responses. We ran a GLMM predicting correct responses produced after exposure to false statements. Fixed effects were statement difficulty (contrast coded; easy = .5 and hard = -.5), experiment phase (contrast coded; pre-feedback = -.5 and post-feedback = .5), and feedback type (no feedback set as the referent condition). Participants and items were included as random intercepts, and statement difficulty as a random slope.

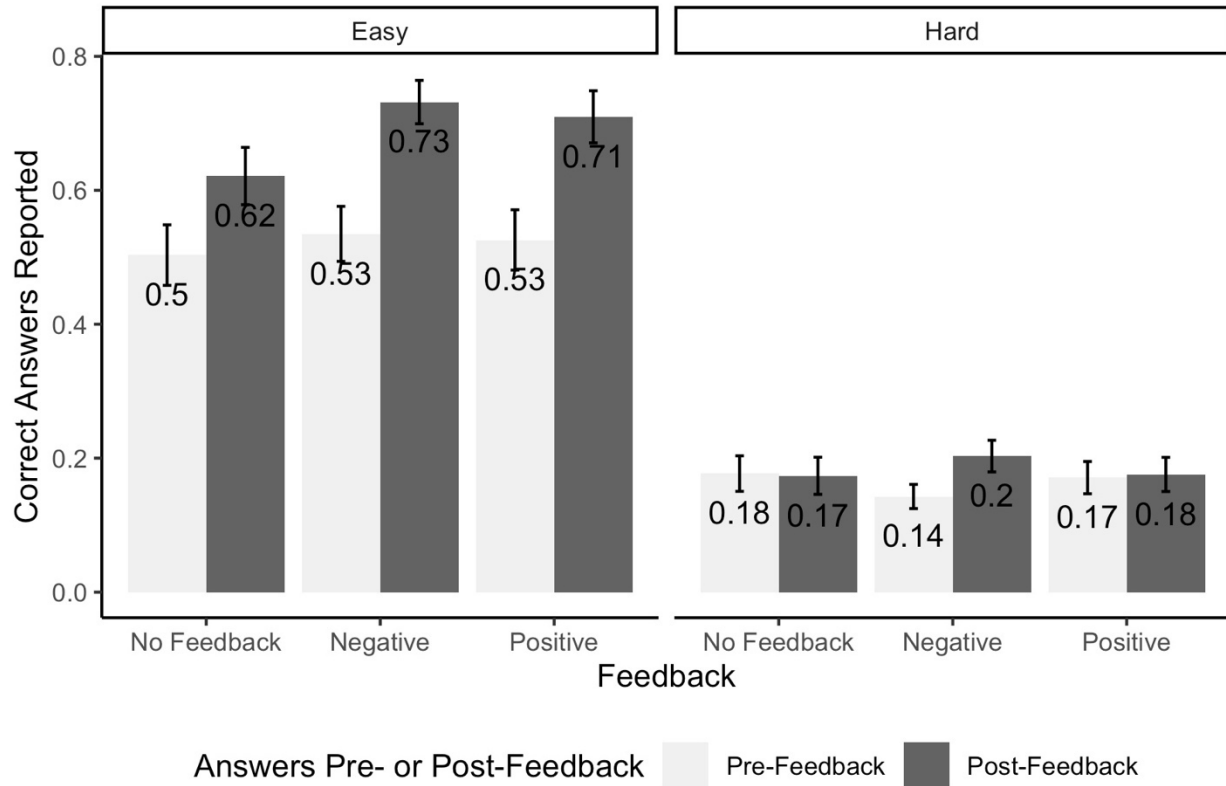
Participants produced more correct responses post-feedback ($M = .44$, $SD = .36$) than pre-feedback ($M = .34$, $SD = .33$), $b = .38$, $z = 3.44$, $p < .001$. There was again a significant experiment phase x statement difficulty interaction, $b = 1.05$, $z = 4.80$, $p < .001$, with simple contrasts revealing that this increase in correct responses occurred more so for questions about easy, $z = 15.14$, $p < .001$, than hard statements, $z = 1.91$, $p = .06$. Although there were no main effects of interim feedback, the predicted experiment phase x feedback type interactions emerged: The increase in correct answers from pre- to post-feedback was greater for participants

who received either negative feedback, $b = .62$, $z = 4.11$, $p < .001$ or positive feedback, $b = .36$, $z = 2.33$, $p = .02$, as compared to participants who received no feedback. Simple contrasts confirmed that in the pre-feedback experimental phase, the three groups showed no difference, $ps > .05$. In the post-feedback phase, only correct answers produced by the positive feedback group differed from those produced by the no feedback group, $z = 2.52$, $p = .03$.

While neither experiment phase x feedback type x difficulty interaction reached significance, $ps > .05$, to clarify the nature of these aforementioned interaction effects, we ran a pairwise contrast to compare the experimental phase x feedback type interaction between easy and hard items. Post-feedback, both positive feedback, $z = 2.55$, $p = .01$, and negative feedback groups, $z = 2.43$, $p = .02$, showed an increase in correct answer production as compared to the no feedback group when statements were easy. However, relative to the no feedback group, only negative feedback resulted in more correct answers pre- to post-feedback when statements were hard, $z = 3.32$, $p < .001$. In other words, positive feedback only increased correct answer productions for easy items, while negative feedback increased correct answer productions for both easy and hard items (see Figure 9).

Figure 9

Correct Response Rates Following False Information in Experiment 4



Note: Correct answer rate split by difficulty of statement at exposure (easy or hard) and feedback type (no feedback, negative feedback, or positive feedback). Light gray bars represent responses made prior to receiving interim feedback and dark gray bars represent responses made after receiving interim feedback. Error bars represent standard error.

Table 4

Mean Rates of Incorrect Lure and Correct Response Production in Experiment 4.

Experiment Phase	Statement Difficulty	Lure	Correct
Control (No Feedback)			
<i>Pre-Feedback Phase</i>			
True	Easy	.00 (.02)	.95 (.06)
	Hard	.03 (.05)	.50 (.27)
False	Easy	.34 (.31)	.50 (.35)

	Hard	.40 (.22)	.18 (.21)
<i>Post-Feedback Phase</i>			
True	Easy	.01 (.03)	.95 (.08)
	Hard	.04 (.05)	.52 (.30)
False	Easy	.27 (.28)	.62 (.33)
	Hard	.39 (.24)	.17 (.22)
Negative Feedback			
<i>Pre-Feedback Phase</i>			
True	Easy	.00 (.01)	.92 (.18)
	Hard	.03 (.04)	.48 (.26)
False	Easy	.29 (.30)	.43 (.33)
	Hard	.36 (.24)	.14 (.14)
<i>Post-Feedback Phase</i>			
True	Easy	.01 (.03)	.92 (.16)
	Hard	.04 (.07)	.44 (.26)
False	Easy	.17 (.22)	.73 (.26)
	Hard	.26 (.24)	.20 (.19)
Positive Feedback			
<i>Pre-Feedback Phase</i>			
True	Easy	.00 (.01)	.92 (.16)
	Hard	.04 (.07)	.46 (.25)
False	Easy	.31 (.29)	.53 (.35)
	Hard	.39 (.22)	.17 (.19)
<i>Post-Feedback Phase</i>			
True	Easy	.00 (.02)	.88 (.21)
	Hard	.03 (.05)	.44 (.27)
False	Easy	.16 (.23)	.71 (.31)
	Hard	.34 (.21)	.18 (.20)

Note: Numbers in parentheses are standard deviations.

Discussion

When participants received feedback concerning their susceptibility to false information, they were less likely to reproduce inaccuracies and more likely to produce correct responses, as compared to participants who did not receive any feedback. Although the benefits of feedback emerged across both hard and easy statements, the effects were more pronounced for easy items,

as participants could more readily leverage their correct prior knowledge to evaluate the inaccuracies (see Figures 8 and 9). These findings offer theoretical insight into the conditions under which people are more or less likely to be influenced by inaccurate ideas, indicating feedback is generally beneficial. Previous work has suggested overconfidence in one's ability to detect and resist false information can discourage evaluation and increase susceptibility to inaccuracies (e.g., Lyons et al., 2021; Salovich & Rapp, 2021). By this account, positive feedback could ironically encourage people to "let their guard down" while reading, while negative feedback should help people recognize their susceptibility and encourage evaluation. But these differential effects were not obtained, with benefits observed for both negative and positive feedback.

One possible reason for this generalized benefit is that feedback surfaced the potential consequences of reading inaccuracies, which people may not have previously considered or acted to remediate. This aligns with the ways in which feedback benefits performance by clarifying the goals and stakes of a task, and in turn motivating the propensity to monitor, recognize, and use different strategies (Hattie & Timperley, 2007; Paris & Winograd, 1990). Also consistent with this idea, people fare better at resisting inaccurate content when they are provided task-specific goals for avoiding falsehoods and are motivated to act on those goals (Andrews-Todd et al., 2021). Similar benefits could emerge when participants are explicitly told that their reliance on potentially inaccurate content is being monitored. Like feedback, this confronts people with the possibility of falling victim to inaccuracies and the need to consider the accuracy of information more critically than they might otherwise. Unlike feedback, it would not require providing people with commentary on their prior performance. We examined this

possibility in Experiment 5 as a means of specifying potential reasons for feedback benefits, and for testing another way of supporting evaluation.

Experiment 5

Experiment 5 examined whether informing people that their susceptibility to false information was being monitored would reduce inaccurate reproductions. Participants again made interest judgments about statements and then answered related questions, this time only for a single set of statements. Prior to judging the statements, participants were either given instructions that mentioned that the task was intended to measure how much they are influenced by false information (disclosure condition) or instructions omitting that information (control condition). All participants were told they would encounter false ideas. We hypothesized that participants told their performance was being monitored would reproduce fewer incorrect lures and produce more correct answers to post-reading questions than would uninformed participants.

Method

Participants

We targeted one-hundred and sixty subjects, 80 per between subject condition. One hundred and sixty-five undergraduates participated in the study in exchange for course credit, none of whom participated in Experiment 4. We again excluded participants who failed a multiple-choice comprehension check or English language check, or reported looking up answers, leaving a total of 157 participants (96 female; M age = 18.95 years) and 12,560 observations.

Design

The experiment followed a 2 (statement validity: true or false) x 2 (statement difficulty: easy or hard) x 2 (instruction condition: disclosure or control) mixed design. Statement validity

and difficulty were manipulated within subjects and instruction condition was manipulated between subjects.

Materials

We used the same 80 facts and open-ended questions as in Experiment 4. The full set of materials can be found on OSF (osf.io/cq673/).

Procedure

After providing informed consent, participants randomly received one of two instructions at the beginning the experiment. Half of the participants received disclosure instructions which included a detailed explanation of the experimental procedure and notification they were being monitored: “In the first part of the study, you will read a series of statements one at a time and indicate how interesting you find each statement to be. Some of the statements are true and some of them are false. Later in the study, you will be asked a series of general knowledge questions. This task will measure how much your answers are influenced by the false information that you read.” The other half of the participants received control instructions, which were the same save for omitting the final sentence of the disclosure paragraph.

Participants then viewed and answered questions about all 80 true and false general knowledge statements, presented one-at-a-time in a unique random order. All participants rated the statements for how interesting they were on a scale from 1 (*very uninteresting*) to 6 (*very interesting*). Immediately after completing the first exposure task, participants were asked to complete a general knowledge questionnaire and to “answer based on what you believe to be true about the world.” They then answered 80 open-ended questions related to the earlier read statements in a unique random order by typing their answers in a textbox. While they were

instructed not to look up any answers, they were told that they could leave the box blank or write unsure/no answer if they did not know the answer.

Results

A pre-registration of planned analyses for Experiment 5 is available on OSF (osf.io/cq673/).

Questionnaire coding

All responses ($N = 12,560$) were coded using the same procedure and criteria as Experiment 4. Inter-rater reliability for dual-coded responses was $\kappa = .94$ with all disagreements resolved through discussion.

Models

Data were analyzed using GLMM models as in Experiment 4, with separate examinations of incorrect lure and correct productions. Model specifications and outputs for Experiment 5 are publicly available at <https://rpubs.com/nsalovich/feedback-exp2>.

Incorrect Lure Productions

Incorrect lure production rates are summarized in Table 2 and Figure 4. To analyze these reproductions, incorrect lure responses were coded as 1 and all other responses as 0. We first examined how statement validity and difficulty affected lure production. The GLMM included statement validity (contrast coded; true = -.5, false = .5) and statement difficulty (contrast coded; easy = -.5 and hard = .5) as fixed effects, participant and item as random intercepts, and statement validity and difficulty as random slopes.

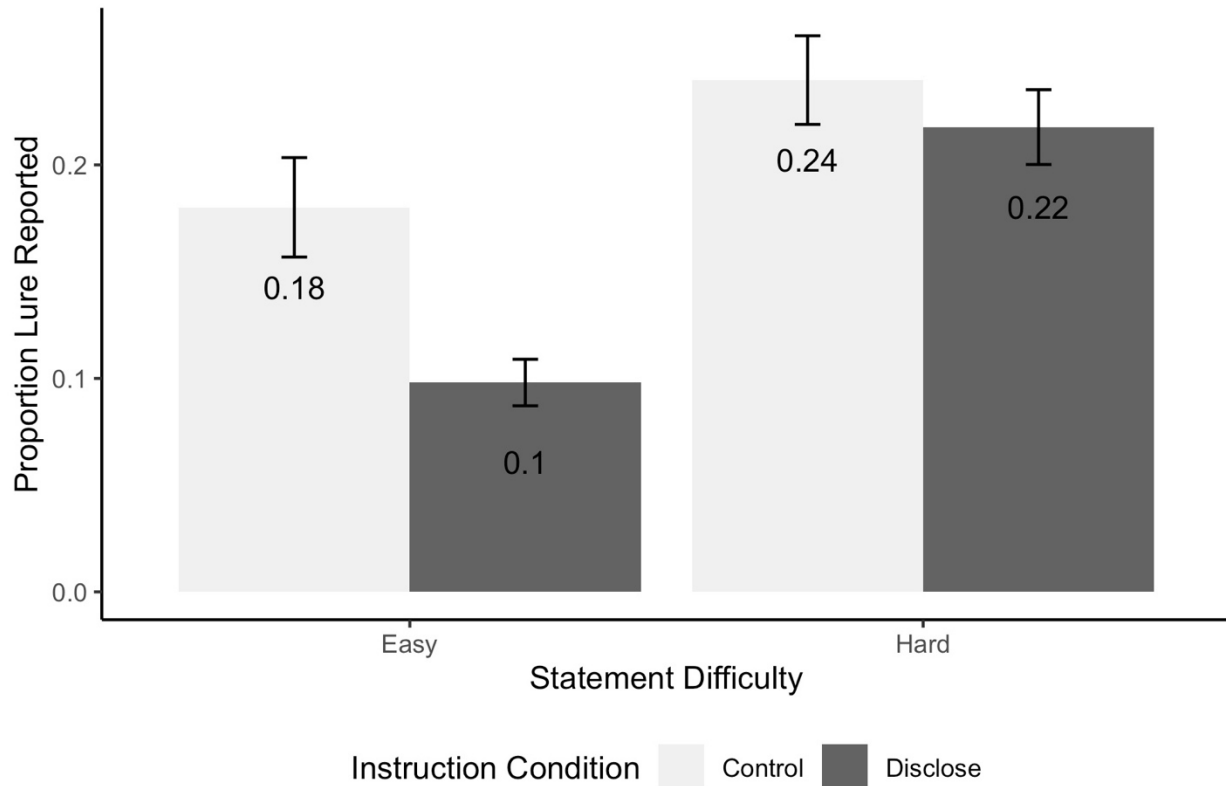
We observed a main effect of difficulty, with participants producing incorrect lures more often for questions about hard ($M = .13$, $SD = .16$) than easy statements ($M = .07$, $SD = .13$), $b = 1.19$, $z = 4.14$, $p < .001$. Participants also reproduced more incorrect lures after reading

statements containing those inaccuracies ($M = .18$, $SD = .17$) as compared to spontaneously producing those incorrect lures after reading true statements ($M = .02$, $SD = .03$), $b = 2.60$, $z = 16.53$, $p < .001$. We also observed a significant statement validity x statement difficulty interaction, $b = -.91$, $z = -3.30$, $p < .001$. Simple contrasts revealed that participants were overall more likely to produce incorrect lures after exposure to false as compared to true statements, although this pattern was more readily observed for questions about hard, $OR = 2.15$, $z = 14.48$, $p < .001$, than easy statements, $OR = 3.05$, $z = 11.97$, $p < .001$. This replicated Experiment 4.

How did task instructions affect reproductions of incorrect lures? To address this question, we ran a GLMM predicting incorrect lure production following exposure specifically to false statements. Statement difficulty (contrast coded; easy = $-.5$ and hard = $.5$) and task instructions (contrast coded; disclosure = $-.5$ and control = $.5$) were entered as fixed effects, participants and items as random intercepts, and statement difficulty as a random slope. There was a marginal main effect of instruction type: Participants reproduced fewer incorrect lures if they had earlier received disclosure instructions ($M = .09$, $SD = .17$) as compared to control instructions ($M = .12$, $SD = .17$), $b = .37$, $z = 1.84$, $p = .07$. A difficulty x instruction type interaction emerged, $b = -.56$, $z = -2.55$, $p = .01$. Simple contrasts revealed that participants who received disclosure instructions reduced inaccurate reproductions for easy, $z = 2.62$, $p = .009$, but not hard items, $z = .42$, $p = .67$. Informing participants that the task monitored their susceptibility to false information reduced incorrect reproductions, but only for statements about which participants were likely to possess prior knowledge.

Figure 10

Error Rates Following False Information in Experiment 5



Note: Error rates following false information split by difficulty of statement at exposure (easy or hard) and instruction type (control or disclosure). Error bars represent standard error.

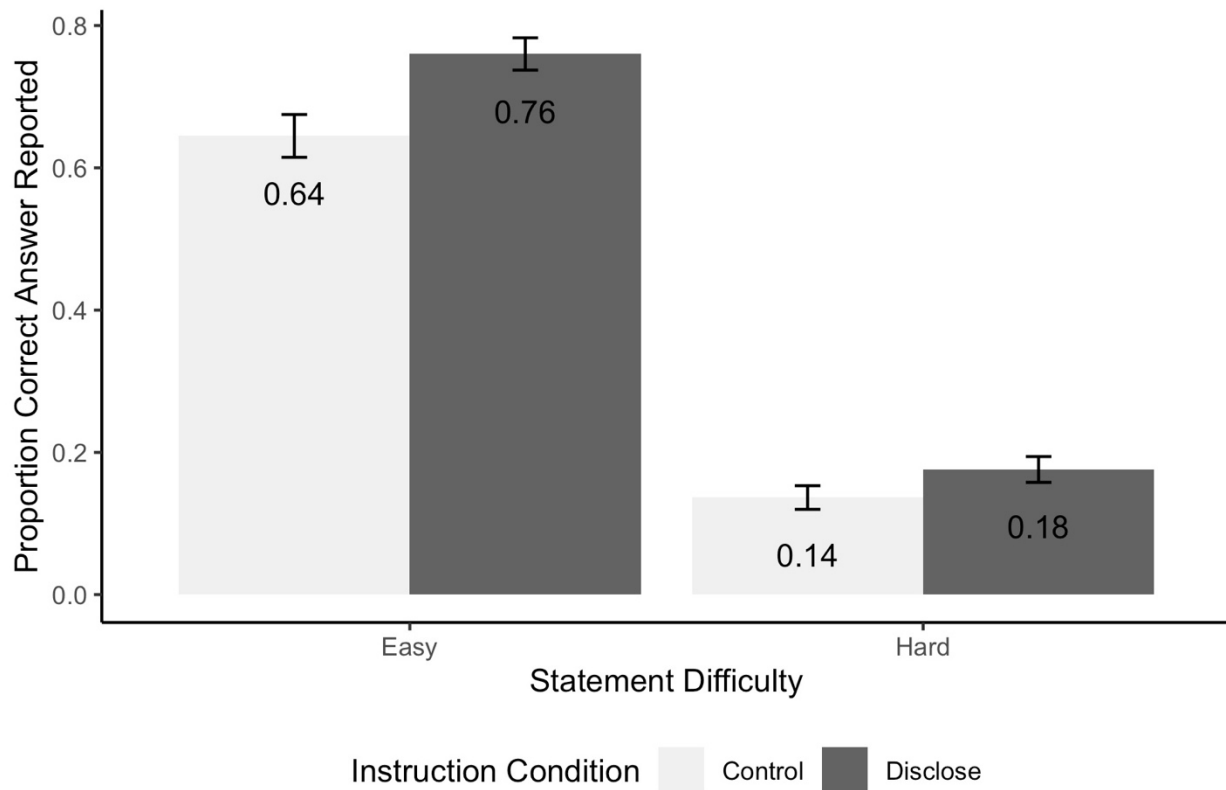
Correct Productions

Correct response rates are summarized in in Table 2 and Figure 5. To analyze correct answers, responses that were correct were coded as 1 and all other responses as 0. We first examined whether statement validity and difficulty affected correct responses. To do so, we ran a GLMM that included statement validity (contrast coded; true = .5 and false = -.5) and difficulty (contrast coded; easy = .5 and hard = -.5) as fixed effects, participant and item as random intercepts, and statement validity and difficulty as random slopes. Participants were more likely to produce correct responses to questions about easy ($M = .80$, $SD = .23$) than hard statements ($M = .26$, $SD = .21$), $b = 3.72$, $z = 16.02$, $p < .001$. They also produced more correct responses after

exposure to true ($M = .62$ $SD = .33$) as compared to false statements ($M = .43$, $SD = .34$), $b = 1.62$, $z = 17.85$, $p < .001$. There was also a significant statement validity x statement difficulty interaction, $b = .27$, $z = 2.16$, $p = .03$. Simple contrasts revealed that there was a greater difference in correct responses following exposure to true and false easy items, $b = 1.76$, $z = 15.28$, $p < .001$, than hard items, $b = 1.49$, $z = 14.05$, $p < .001$.

We next assessed how instructions affected correct responses. We ran a GLMM predicting correct responses following exposure to false statements by statement difficulty (contrast coded; easy = .5 and hard = -.5) and task instructions (contrast coded; disclosure = .5 and control = -.5). Participants and items were included as random intercepts, and statement difficulty as a random slope. A significant main effect of instruction type indicated that participants who received disclosure instructions overall produced more correct answers ($M = .55$ $SD = .35$) than did participants who received control instructions ($M = .50$ $SD = .34$), $b = .57$, $z = 2.77$, $p = .006$. The predicted statement difficulty x instruction type interaction did not reach significance, $b = .28$, $z = 1.35$, $p = .18$. However, given the pattern of results was in the predicted direction of our registered analyses, we investigated differences in instruction type on easy and hard items using simple contrasts. Consistent with our hypothesis, for easy items, participants who received disclosure instructions produced more correct responses than did participants who received control instructions, $z = 3.07$, $p = .002$; that difference was marginal for hard statements, $z = 1.88$, $p = .06$. Informing participants that their susceptibility to false information was being monitored increased correct responses, especially when participants likely possessed accurate prior knowledge about those ideas.

Figure 11

Correct Response Rates Following False Information in Experiment 5

Note: Correct response rates following false information split by difficulty of statement at exposure (easy or hard) and instruction type (control or disclosure). Error bars represent standard error.

Table 5*Mean Rates of Incorrect Lure and Correct Response Production in Experiment 5*

Statement Validity	Statement Difficulty	Lure	Correct
<i>Control Instructions</i>			
True	Easy	.01 (.02)	.88 (.18)
	Hard	.04 (.03)	.35 (.20)
False	Easy	.18 (.21)	.64 (.27)
	Hard	.24 (.18)	.14 (.15)

Disclosure Instructions

True	Easy	.00 (.02)	.91 (.15)
	Hard	.03 (.04)	.36 (.24)
False	Easy	.10 (.10)	.76 (.20)
	Hard	.22 (.26)	.18 (.16)

Note: Numbers in parentheses are standard deviations.

Discussion

Participants who were informed their susceptibility to false information was being monitored reproduced fewer inaccuracies and produced more correct answers than did participants who were only informed of the task procedure. This difference obtained despite all participants being explicitly told they would read false information. Instructions that emphasized performance monitoring benefited responses as consistent with feedback and with evaluation.

General Discussion

People's use of false information is reduced when they are confronted with their potential susceptibility to inaccuracies. In Experiment 4, participants who received feedback on their task performance, whether positive (highly resistant to inaccuracies) or negative (highly influenced by inaccuracies), reproduced fewer false ideas and more correct answers on a second iteration of the task, as compared to participants who did not receive feedback. In Experiment 5, similar benefits were obtained when participants were simply informed that their susceptibility to false information was being monitored during the task. Rates of inaccurate reproductions in the feedback (Experiment 4) and monitoring (Experiment 5) conditions were comparable to previous studies in which participants were explicitly told to evaluate information (Salovich, Kirsch, et al., 2022), despite the current tasks explicitly requiring non-evaluative (i.e., interest) judgments.

Previous work has documented how overconfidence in one's resistance to inaccuracies can discourage evaluation (Lyons et al., 2021; Salovich & Rapp, 2021). Making people aware of

this issue and/or their propensities towards relying on inaccurate information might therefore support performance, with feedback one way of encouraging awareness of confidence-ability discrepancies. The negative feedback in Experiment 4 provided such a prompt, suggesting the need for engaging in evaluation. But benefits were also observed following positive feedback, wherein participants were told they were highly resistant to inaccuracies. This is perhaps surprising as indicators of successful performance seem likely to increase rather than decrease confidence. One possibility for the benefit of positive feedback, and for feedback in general, is it may have afforded the opportunity to contemplate the consequences of exposures to inaccurate ideas — a risk people might not routinely consider that is both relevant and useful to the task. This is consistent with the general finding that feedback improves awareness, assessment, and adaption of in-the-moment task strategies and goals (Lee et al., 2010; Nietfield, et al., 2006; Roll et al., 2011; Schraw & Moshman, 1995).

In accord with this possibility, in Experiment 5, informing participants that their susceptibility to inaccuracies was being monitored also reduced inaccurate reproductions and increased correct responses. Participants in the feedback conditions in Experiment 4, and in the monitoring condition in Experiment 5, received information that drew attention to aspects of performance (i.e., their susceptibility; the utility of considering information accuracy; the benefits of evaluation) relevant for noticing and acting on inaccurate content, encouraging or even amplifying awareness of epistemic strategies for success on the task (Hattie & Timperley, 2007; Kazdin, 1974). Belief that one is being evaluated can also motivate productive learning strategies (Bond et al, 1983; Cottrell, 1972), including the detection and rejection of inaccurate ideas (Andrews-Todd et al., 2021). These influences likely worked in concert to encourage an

accuracy focus in Experiments 4 and 5, without explicitly instructing participants to maintain an evaluative mindset.

These experiments also foreground that people's thoughts and beliefs about their resistance to inaccuracies play a crucial role with respect to the influence of false statements. Many studies have documented that people seem unmotivated to engage in evaluation despite such behaviors being necessary and useful (e.g., Bago et al., 2020; Pennycook, Epstein, et al., 2021; Pennycook & Rand, 2022). For example, explicit warnings about inaccurate content do not easily encourage evaluation (e.g., Ecker et al., 2010; Marsh & Fazio, 2006), at times necessitating additional motivational instructions to obtain benefits (e.g., Andrews-Todd et al., 2021; Donovan & Rapp, 2020; Marsh & Fazio, 2006). This occurred here as well: Participants in Experiment 4 who did not receive feedback, and participants in Experiment 5 who were told they would read false statements but not that they were being monitored, revealed traditionally reported reproductions of inaccurate ideas. Changes in these patterns required providing performance feedback or warnings that performance was being monitored. These two manipulations tap into people's beliefs about their susceptibility to false information, and thoughts about whether they should engage in evaluation. They also identify powerful motivators for encouraging critical engagement with information.

The results therefore have practical implications for combating endorsements of, belief in, and the spread of inaccurate information. People may not opt to consider the accuracy of what they read or hear, which is concerning given the amount of information provided in news reports, social media feeds, personal narratives, fiction, gossip, and from other sources and outlets. They can engage with these materials in ways that ignore content validity, instead focusing on personal interest, humor, novelty, relatability, and suspense. One tactic to combat the influence

of inaccurate information could involve confronting individuals with their susceptibility to inaccuracies, either through direct feedback or indirectly through monitoring. People's thoughts about their susceptibility to misleading or false ideas could be added to existing initiatives, which often focus more on explicit instructions and prompts to dutifully evaluate content. For example, fact-checking and accuracy nudges have obtained promising benefits for reducing belief in falsehoods, at least in the short-term (e.g., Carey et al., 2022; Roozenbeek et al., 2021). Fact-checks might remind people of their history of engaging with false content (e.g., article X that you liked/shared/clicked on is false), offering a form of performance feedback. Likewise, nudges that remind people that their interactions with inaccurate content are being monitored, including notifications that one has engaged with fake news, or badges awarded for sharing high-quality information or sources, could encourage more persistent accuracy considerations. Of course, there may be important, unintended consequences of feedback and monitoring, including concerns about the privacy and security of one's information and behaviors. Work is needed to assess the ways in which potential interventions, whether constituting explicit instructions or tools that monitor and offer personalized feedback, affect engagement and comfort with content and dissemination systems.

Despite concerns over the circulation of false content, people regularly identify accuracy as crucially important for the information they engage with and their own information-based reputations (e.g., Altay et al., 2020; Knight Foundation, 2018; Pennycook, Epstein, et al., 2021; Waruwu et al., 2020). The current study demonstrates how leveraging people's beliefs about their susceptibility and attention to inaccuracies (i.e., through feedback and monitoring) can support evaluation, decreasing well-replicated patterns of inaccurate reproductions, and increasing use of accurate knowledge.

Given the clear applications of the findings obtained in the experiments thus far, Chapter 3 aims to both apply and expand these results to information resembling naturalistic experiences in social media environments. In Experiment 6, we examine whether people's general knowledge can be problematically influenced by reading false information within tweets. In Experiment 7, we explore how both evaluative and non-evaluative tasks and mindsets affect people's susceptibility to inaccuracies on social media.

Chapter 3:

The Effects of Evaluative and Non-Evaluative Mindsets on the Influence of Inaccurate Tweets

Exposure to inaccurate information can affect people's everyday judgments in ways that can be life changing, and even at times life threatening. Major world events have been influenced and guided by prevalent false information; for example, some individuals have attributed the results of the 2016 U.S. election and health concerns such as the 2019 measles outbreak to the circulation of inaccurate claims and ideas (Allcott & Gentzkow, 2017; Carrieri et al., 2019; Dewey, 2016; Parkinson, 2016). Although people encounter false and misleading ideas in a variety of modalities (e.g., Pennycook & Rand, 2021), there is substantial specific concern about falsehoods running rampant on social media platforms (Guess et al., 2019; Lazer et al., 2018; Pasquetto et al., 2020; Thebault, 2019; Vosoughi et al., 2018). With their approachable design, social appeal, and ready availability, apps like Facebook and Twitter are the preferred news source for many people. However, the dynamic, open nature of the platforms makes them an obvious locus for fake news and inaccuracies. In response to these concerns, citizens, policy makers, and even social media companies have expressed a desire to curb the amount of inaccurate information presented and shared on these sites.

Fortunately, recent studies (including in this dissertation) suggest that contemplating the accuracy of presented information may be useful for protecting against the influence of false or misleading claims (Brashier et al., 2020; Rapp, Hinze, et al., 2014; Pennycook & Rand, 2022; Salovich et al., 2021). While most studies have involved lab-based manipulations, some researchers have begun to apply these findings to real-world contexts, including in online environments. For example, work has considered whether an accuracy focus can dissuade the

sharing of false political news stories (Bago et al., 2020; Calvillo & Smelter, 2021; Pennycook, Epstein, et al., 2021; Pennycook et al., 2020; Roozenbeek et al., 2021). A critical assumption motivating these sorts of interventions is that people often fail to consider the accuracy of information they encounter online. Thus, drawing people's attention to accuracy when deciding what to believe or share could help people recognize problematic content they may have otherwise overlooked, and help reduce dissemination of falsehoods (e.g., Pennycook, Epstein, et al., 2021; Pennycook & Rand, 2022).

Previous research has centered around assessing the downstream effects of an accuracy focus, but less work has examined the everyday tasks and mindsets that could encourage or dissuade such a focus during experiences with inaccurate information. It also remains unclear, given the lack of baseline comparisons, whether the "control" conditions that are often contrasted with an accuracy focus might actually *reduce* evaluative tendencies. If so, the reported patterns could reflect less in the way of accuracy enhancements, and more so attenuated evaluation when instructions do not encourage attention to accuracy at all. For example, certain non-evaluative tasks and mindsets used as active controls in previous studies (e.g., judgments of interest or humor; e.g., Brashier et al., 2020; Calvillo & Smelter, 2021; Pennycook, Epstein, et al., 2021; Salovich, Kirsch, et al., 2022) could encourage people to think *less* critically about the accuracy of information than might occur during many other kinds of comprehension experiences. This is especially relevant to the ways in which people approach information on social media, as they may be focused on non-evaluative considerations of the information, including but not limited to interest, humor, novelty, relatability, and suspense, rather than the accuracy of the content. Given these issues, this chapter aims to characterize how the activities that participants engage in

during exposure to information in a social media context might differentially affect evaluative processing, and therefore downstream reliance on false ideas.

To conduct this examination, we adapted the statements used in the experiments from Chapters 1 and 2 into tweets containing either correct or incorrect general knowledge information. The purpose of developing these tweet stimuli is twofold: First, as compared to the isolated true and false statements from Chapters 1 and 2, tweets help offer a realistic way to present information on various topics, which increases the external validity of this work. Social media posts allow for true and false information to be embedded in actual discourse framings, utilizing the syntactic, pragmatic, and contextual wrappings that are often encountered during everyday online communication. Twitter in particular is a platform in which users regularly encounter distinct snippets of information from various sources, and has become an increasingly popular way for people to inform and update their understandings of the world (Walker & Matsa, 2021).⁶ Representing everyday discourse experiences is crucial to our goal of assessing and generalizing the downstream effects of evaluative and non-evaluative tasks and mindsets, as the mode and context in which information is presented can affect both the influence of inaccuracies, as well as how easily falsehoods might be detected and corrected (e.g., Corneille et al., 2020; Baker & Wagner, 1987; Fazio, Dolan, et al., 2015). An additional reason to use the tweet stimuli is that they allow us to investigate and contrast the downstream effects of decisions specific to social media posts. The epistemic considerations involved with judgments such as whether to “like” or “share” a tweet are unique to information presented on social media; therefore, asking

⁶ To reflect more naturalistic experiences, some researchers have presented factual inaccuracies in fictional stories, such as through conversations between characters (e.g., mentioning someone’s travel plans to Oslo as the capital of Finland, when the capital is really Helsinki; Marsh et al., 2003). This method, while informative, opens up other issues related to the distinctive ways that people can approach information offered in fiction versus from other sources (Prentice et al., 1997).

participants to adopt social media-specific tasks and mindsets necessitates stimuli that represent the kind of information they would view on those platforms.

Using newly-developed stimuli, Chapter 3 includes two experiments examining how the tasks and mindsets with which people approach information on social media can affect the degree to which they are susceptible to false information. In Experiment 6, we provide an initial test of these new materials, examining whether reading tweets containing true or false information affects people's subsequent responses to general knowledge questions. In Experiment 7, we investigated how judging tweets based on how accurate they are (accuracy mindset), how interesting they are (interest mindset), and how likely someone would be to "like" the tweet on social media (social media mindset) differentially affected people's reliance on presented inaccuracies.

These two experiments allowed a crucial test of whether an accuracy focus during exposure to true and false tweets can reduce the reproductions of inaccuracies and increase productions of correct answers following exposures to false information. While other studies have considered how an accuracy focus can reduce sharing of false or misleading ideas (mostly with attention at least to date to political news headlines), no study to date has investigated the impact of evaluation on people's actual *reproductions* of false information seen on social media. We believe this is a crucial contribution to extant literature because it provides a more direct measure of whether evaluation can dissuade belief in inaccuracies as seen on social media, and tests whether an accuracy focus can encourage people to rely on their correct prior knowledge rather than the encountered inaccuracies (Rapp & Salovich, 2018). We also were able to examine how the common social media task of liking a post might affect participants' subsequent use of inaccuracies presented in tweets, and how liking might compare to performance after judging

tweets for accuracy or interest. For further insight into the activities and considerations people engage in during these judgment tasks, in Experiment 7, we also analyzed response times, the relationship between initial ratings of interest, accuracy, or whether to “like” the tweet and participants’ reproductions of presented information, and qualitative testimonies from participants on their accuracy contemplations. Together, the experiments in Chapter 3 better elucidate the consequences of evaluative and non-evaluative mindsets when considering information on social media, and offer practical implications for combatting the belief in and spread of misinformation online.

Experiment 6

In Experiment 6 we validated the new Twitter stimuli, testing whether viewing true and false tweets affected people’s answers to general knowledge questions. Participants read and made non-evaluative interest judgments about a series of tweets containing true information (e.g., “Finally in Paris! Can’t wait to explore the capital of France”), false information (e.g., “Finally in Marseille! Can’t wait to explore the capital of France”), or uninformative filler tweets (e.g., “Finally landed! Can’t wait to explore the capital of France”). Half of the tweets contained easy information, meaning that they were likely well-known to participants (e.g., “Happy Thanksgiving to all! It’s been over 400 years since the pilgrims sailed to America on the [Mayflower/Godspeed]”) and half contained hard information, meaning they were less likely to be known to participants (e.g., “Walked across the George Washington Bridge over the [Hudson/Mississippi]. Feeling very historic today”). Afterwards participants were asked open-ended questions corresponding to each of the previously presented statements (e.g., “What is the capital of France?”). We were specifically interested in reproductions of incorrect lures from the statements (e.g., “Marseille”) as well as correct responses (e.g., “Paris”).

In line with previous work, we predicted that participants would reproduce more incorrect lures after exposure to tweets containing false as compared to true or filler information, particularly when the information was lesser-known or “hard.” We also predicted that participants would produce fewer correct answers after exposure to tweets containing false as compared to true or filler information, particularly when the information was more well-known or “easy.”

Method

The pre-registration for this experiment can be found at https://aspredicted.org/CB8_4Z9.

Participants

The sample included 90 Amazon Mechanical Turk workers (28 females, 60 males, 2 preferred not to say; M age = 37.43 years) to match the number of observations ($n = 6,480$) in Salovich, Kirsch, et al., (2022) Experiment 1. Given the plan to analyze data using mixed effect models, this allowed sufficient observations of each item type both within and across participants to achieve adequate power while reducing Type 1 Error rates (Luke, 2017). We excluded participants who failed a comprehension check, an English language check, or reported looking up answers, and continued recruitment until the pre-registered target sample size was reached. All participants were paid equal to or above the U.S. minimum wage at the time of data collection.

Participants on average reported using 3.67 different social media apps with YouTube (81%) and Facebook (68%) being most popular platforms. Over half (58%) of participants reported using Twitter. Most (58%) respondents reported using social media between 10-30 minutes per day (18%), 31-60 minutes per day (21%), or 1-2 hours per day (19%).

Design

The experiment followed a 3 (tweet validity: true, false, or filler) x 2 (tweet difficulty: easy or hard) design. Both tweet validity and difficulty were manipulated within-subjects. Participants were asked knowledge questions related to all 72 tweets, which allowed us to measure rates of incorrect lure and correct answer responses.

Materials

Seventy-two tweet-like stimuli were adapted from the general knowledge statements used in Experiments 1-5. For example, the statement “The capital of France is [Paris/Marseille]” was changed to “Finally in [Paris/Marseille]! So cool to explore the capital of France” and presented in a realistic, tweet-like format⁷ (see Table 6 and Appendix B for examples). Half ($n = 36$) of the tweets referenced well-known information or “easy” items, and the other half of the tweets referenced lesser-known information or “hard” items. We also created uninformative, neutral filler versions of each of the tweets, syntactically similar to the true and false tweets, but omitting the key content to be tested later in the experiment (e.g., “Finally landed! So cool to explore the capital of France”). Thus, each tweet had three possible versions: true, false, and neutral filler (see Table 6). At test, all participants responded to open-ended questions related to the 72 tweets (e.g., “What is the capital of France?”). The complete set of Twitter stimuli are publicly available on OSF (osf.io/wh4su/).

Table 6

Example Statement and Converted Twitter Stimuli

Item Type and Difficulty	Statement/Tweet Validity
--------------------------	--------------------------

⁷ Names and usernames were randomly selected via an online American/English name generator. Each tweet (e.g., item 1, item 2, etc.) was presented with a different name and username, but names were kept consistent within versions of the same item (e.g., true, false, filler versions of item 1).

Example Easy Item (#19)

Statement	True	The capital of France is Paris.
	False	The capital of France is Marseille.
	Filler	The capital of Canada is Ottawa.
Tweet	True	Finally in Paris! Can't wait to explore the capital of France.
	False	Finally in Marseille! Can't wait to explore the capital of France.
	Filler	Finally landed! Can't wait to explore the capital of France.

Example Hard Item (#48)

Easy	True	Michelangelo's statue of David is located in Rome.
	False	Michelangelo's statue of David is located in Florence.
	Filler	The Amalfi Coast is located in southern Greece.
Hard	True	In Rome for the weekend and making time to see Michelangelo's statue of David in person.
	False	In Florence for the weekend and making time to see Michelangelo's statue of David in person.
	Filler	In Italy for the weekend and making time to see Michelangelo's statue of David in person.

Procedure

After providing informed consent, participants completed the *exposure phase*. They were presented the following instructions: “In this next part of the study, we’d like you to pretend that you are reading the following tweets as you are scrolling through your Twitter timeline. Please answer the questions in this survey to the best of your ability without using any external resources to help you, such as looking up information online.” Participants rated each tweet on how interesting they found it to be, on a scale from 1 (*uninteresting*) to 5 (*interesting*) similar to past projects (Brashier et al., 2020; Calvillo & Smelter, 2020; Salovich, Kirsch, et al., 2022). Tweets were presented one-at-a-time in a unique random order for each participant. Each participant viewed 36 easy and 36 hard tweets, with one third of the statements true, one third false, and one third neutral fillers.

Immediately after completing the exposure phase, participants completed the *test phase*. Participants were asked to complete a general knowledge questionnaire, and to “answer based on what you believe to be true about the world.” They then answered 72 open-ended questions related to the earlier read tweets in a unique random order by typing their answers in a textbox. They were instructed not to look up any answers, but had the option to leave the textbox blank or write unsure/no answer if they did not know the answer.

In addition to basic demographic questions (e.g., gender, age, race), we also asked about participants’ social media use (e.g., “In the past week, on average, approximately how much time per day have you spent actively using social media?”; Ernala et al., 2020) and on which social media platforms they consider themselves to be active users (e.g., Twitter, Facebook, Snapchat, etc.).

Results

Questionnaire coding

Responses ($N = 6,480$) were coded as correct, incorrect lure, incorrect other, or unsure/blank. Two raters independently coded a quarter of the responses in the data set, with the remaining coded by one rater only. Inter-rater reliability for dual-coded responses was reliably high ($\kappa = .92$), with all disagreements resolved through discussion.

Models

Data analyses were run using generalized linear mixed effect models (GLMM) with the R packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2014), with separate examination of incorrect lure and correct answer productions. Simple contrasts were calculated using the R package emmeans (Lenth, 2019). Model specifications and outputs for Experiment 6 are publicly available at <https://rpubs.com/nsalovich/twitter-exp1>.

As in Chapters 1 and 2, we conducted separate analyses of incorrect lure and correct answer responses to test our hypothesized effects, which is consistent with previous research examining how presentations of information affect people's responses to open-answer questions (e.g., Donovan & Rapp, 2020; Kelley & Lindsay, 1993; Fazio et al., 2013; Fazio & Marsh, 2008; Marsh & Fazio, 2006; Marsh et al., 2003). Tables summarizing the distribution of response types for both Experiments 6 and 7 are available as supplemental material on OSF (osf.io/wh4su/).

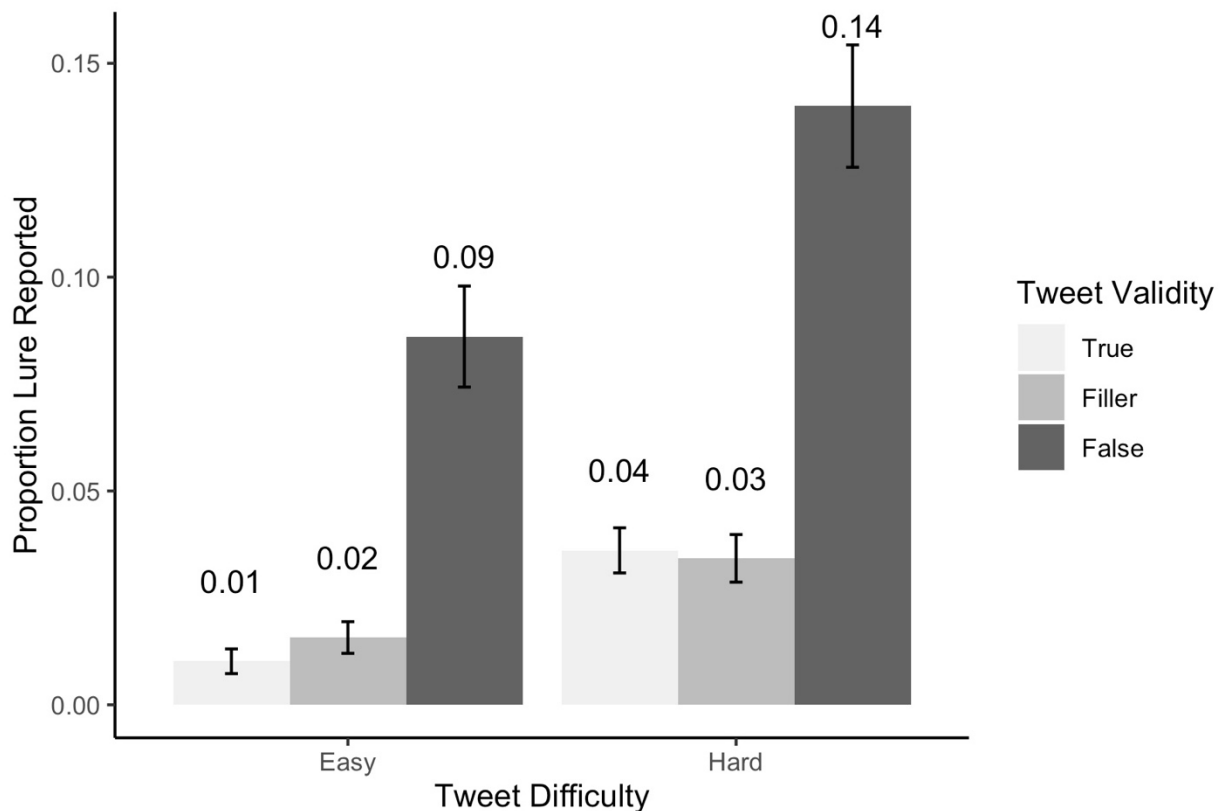
Incorrect Lure Productions

Incorrect lure reproduction rates are summarized in Table 7 and Figure 12. To analyze these reproductions, incorrect lure responses were coded as 1 and all other responses as 0. We first examined how tweet validity and difficulty affected incorrect lure reproduction. The GLMM included tweet validity (false tweets set as referent condition) and tweet difficulty (contrast coded; easy = -.5 and hard = .5) as fixed effects, participant and item as random intercepts, and tweet validity and difficulty as random slopes.

As predicted, participants reported more incorrect lures after exposure to tweets containing that false information ($M = .11$, $SD = .13$) as compared to tweets that contained true information ($M = .02$, $SD = .04$), $b = -1.76$, $z = .24$, $p < .001$, or no relevant (“filler”) information ($M = .03$, $SD = .05$), $b = -1.70$, $z = -6.63$, $p < .001$. While participants produced overall more incorrect lures to answer hard ($M = .07$, $SD = .10$) versus easy questions ($M = .04$, $SD = .08$), $b = .91$, $z = 2.12$, $p = .03$, no tweet validity \times tweet difficulty interactions emerged, $ps > .05$. People reproduced incorrect lures from tweets containing both well-known and lesser-known false information.

Figure 12

Error Rates Following False Information in Experiment 6

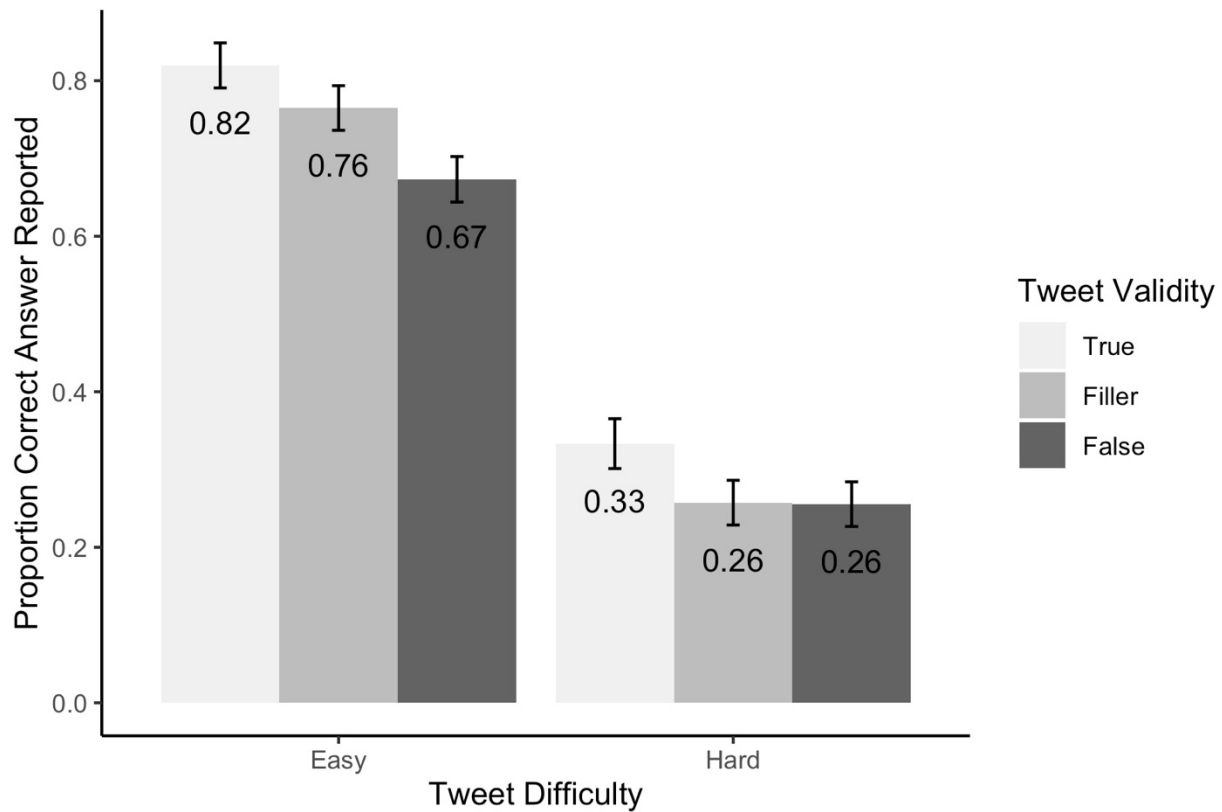


Note: Error rates split by difficulty (easy or hard) and validity (true, filler, or false) of tweet at exposure. Error bars represent standard error.

Correct Answer Productions

Correct response rates are summarized in in Table 7 and Figure 13. To analyze correct answers, correct responses were coded as 1 and all other responses as 0. To examine whether tweet validity and difficulty affect correct responses, we ran a GLMM including tweet validity (false tweets set as referent condition) and difficulty (contrast coded; easy = .5 and hard = -.5) as fixed effects, participant and item as random intercepts, and tweet validity and difficulty as random slopes.

Participants reported fewer correct answers after exposure to tweets containing false information ($M = .46$, $SD = .34$) as compared to tweets that contained true information ($M = .58$, $SD = .38$), $b = .95$, $z = 8.00$, $p < .001$, or no relevant (“filler”) information ($M = .51$, $SD = .37$), $b = .37$, $z = 4.02$, $p < .001$. They also produced overall produced fewer correct answers to hard ($M = .28$, $SD = .28$) as compared to easy questions ($M = .75$, $SD = .28$), $b = 2.83$, $z = 11.33$, $p < .001$. These main effects were qualified by significant tweet validity x tweet difficulty interactions (false vs true: $b = .93$, $z = 4.47$, $p < .001$; false vs filler: $b = .80$, $z = 4.02$, $p < .001$). Simple contrasts revealed that correct responses were less likely after exposure to false versus true tweets more so when information was easy, $z = 8.62$, $p < .001$, as compared to hard $z = 3.18$, $p = .004$. People also produced fewer correct answers following exposure to false versus filler tweets when the information was easy $z = 5.80$, $p < .001$, with no difference for hard information, $z = 4.02$, $p = .97$. Overall, exposure to false information in tweets reduced correct answers to the general knowledge questions, with greater differences emerging for familiar versus unfamiliar information.

Figure 13*Correct Response Rates Following False Information in Experiment 6*

Note: Correct response rates split by difficulty (easy or hard) and validity (true, filler, or false) of tweet at exposure. Error bars represent standard error.

Table 7*Mean Rates of Incorrect Lure and Correct Response Production in Experiment 6*

Tweet Validity	Tweet Difficulty	Lure	Correct
True	Easy	.01 (.03)	.82 (.27)
	Hard	.04 (.05)	.33 (.30)
False	Easy	.09 (.11)	.67 (.28)
	Hard	.14 (.14)	.26 (.27)

Filler	Easy	.02 (.04)	.76 (.27)
	Hard	.03 (.05)	.26 (.27)

Note: Numbers in parentheses are standard deviations.

Discussion

In line with previous work, exposures to tweets containing incorrect information increased the rate at which inaccuracies were reproduced by participants as compared to viewing tweets containing true information or no relevant information. These problematic effects emerged for both well-known and lesser-known inaccuracies. Exposure to false tweets also reduced correct answer production relative to viewing tweets containing true information or no relevant information, particularly when the information was familiar to participants. Together these results demonstrate that exposure to tweets containing inaccuracies can result in a problematic influence of those false ideas over accurate prior knowledge.

An interesting observation is that the effect sizes observed in the current study were smaller than in previous work examining the effects of exposure to the same true and false facts in statement form (e.g., see beta values of experiments summarized in Chapter 1). Put differently, there is a greater likelihood that participants will answer with “Marseille” to the question “What is the capital of France?” after exposure to the information as a statement (i.e., The capital of France is Marseille) as compared to exposure to the information embedded in a tweet (i.e., Finally in Marseille! Can’t wait to explore the capital of France). This difference could have emerged for a variety of reasons. For example, one possibility is that information presented as declarative statements may be more likely to be presumed as true due to default assumptions about the accuracy of presented ideas (see Brashier & Marsh 2020; Gilbert et al.1990; Gilbert, 1991). According to the maxim of quality, people assume that conversational partners present information as truthful and do not provide information that is false or not

supported by evidence (Begg et al., 1992; Grice, 1975). It is possible that people may be more likely to unquestioningly accept ideas as valid when they are explicitly stated as facts versus when they are more indirectly communicated in the context of tweets. This is particularly relevant given that the tweets were presented as coming from a random source (i.e., Twitter user) versus unsourced and/or from a trustworthy source, such as directly from the experimenter (e.g., Groggel et al., 2018; Mena et al., 2020; Zawadzka et al., 2016, but see Andrews & Rapp, 2015 wherein people defaulted to trusting information from random partners when no knowledge as to their competence was available). Recent work has also considered potential differences in the influence of falsehoods as a function of the context or environment in which the information is encountered. For example, while repetition traditionally increases the perceived accuracy of information (e.g., Brashier et al., 2020; Dechêne et al., 2010; Fazio et al., 2019; Hasher et al., 1977), Corneille et al. (2020) found repetition *decreased* feelings of trustworthiness when information was presented in a social media context. This demonstrates potential differences in the uptake and reliance of information presented on social media as compared to in an unspecified or ambiguous contexts.

Another possible explanation for the smaller effects observed in the current study could be driven by differences in the memorial traces afforded by statements versus tweet-like presentations. Some processing accounts explain the influence of obvious falsehoods on people's judgments as due to competition and/or interference between recently encoded information and prior knowledge (Lewis & Anderson, 1976; Rapp, Hinze, et al., 2014; Rapp, Jacobina, et al., 2014; Salovich, Kirsch, et al., 2022; Weil et al., 2020). For example, when people read inaccurate information (e.g., "The capital of France is Marseille"), they may encode incorrect associations between concepts (e.g., "Marseille" and "France") without sufficient activation of or

even alongside accurate associations (e.g., “Paris” and “France”). If the inaccurate concept (e.g., “Marseille”) is more familiar or shares sufficient semantic associates with background knowledge, it may retain greater availability than existing accurate understandings (e.g., “Paris;” Anderson, 1981; Rapp, Jacovina, et al., 2014; Storm, 2011), affording use of the false answer over correct knowledge on later tasks. Statements may more directly activate and link the sentence components (e.g., “Marseille” and “France”) than tweets, which rely more so on pragmatic elements like bridging inferences to make the connection between the inaccurate answer and core concept communicated by the tweet (Clark, 1975). Consider the example, “Finally in Marseille! Can’t wait to explore the capital of France.” This tweet relies on the reader to infer that the capital of France is Marseille, without such information being explicitly stated by the author. As such, people may experience richer encoding and representations of inaccuracies presented in statements than in tweets, leading to higher rates of problematic retrieval of those false ideas (e.g., Fazio, Dolan, et al., 2015; Salovich, Kirsch, et al., 2022).

Given the goal of understanding and extending the consequences of exposures to inaccuracies as experienced in everyday settings, we believe it is imperative to acknowledge any potential differences in people’s reliance on falsehoods to draw informed theoretical and practical conclusions. That said, the purpose of this dissertation chapter is not necessarily to compare the influence of false ideas based solely on the format of communication, but rather to examine whether and how tasks and mindsets during exposures to inaccuracies on social media might influence subsequent reliance on the ideas. The results of Experiment 6 indicate that reading tweets containing false information elicits problematic downstream consequences on people’s real-world judgments. In Experiment 7, we manipulated the task and mindset that participants engaged in during exposure to the true and false tweets. Specifically, we examined

whether accuracy, interest, and social media judgments differentially affected the degree to which people were influenced by the inaccuracies presented in tweets when answering subsequent questions.

Experiment 7

Previous work examining the utility of evaluative instructions has contrasted accuracy judgments with interest judgments, under the assumption that interest judgments offer a method of encouraging comprehension without prompting consideration of a statement's accuracy. One intriguing and underdiscussed idea is that interest judgments could lead participants to think *less* critically about the accuracy of information than might occur during many kinds of comprehension experiences. This is particularly relevant when considering the goals enacted when people engage with information in online contexts, such as social media, as people may prioritize non-evaluative considerations like interest, humor, or affinity over accuracy. We would expect that the activities that people engage in during exposures to information differentially affects their accuracy considerations, and as such, their downstream reliance on false ideas.

To explore this idea, in Experiment 7 we evaluated how tasks beyond interest and accuracy judgments affect participants' susceptibility to inaccuracies presented in a social media context. As in Experiment 6, participants read a series of true and false tweets one-at-a-time. This time, however, participants were randomly assigned to complete one of three tasks at exposure: Judging the accuracy of each tweet (accuracy mindset condition), judging how interesting each tweet is (interest mindset condition), and judging whether they would 'like' the tweet if they had seen it on social media (social media mindset condition). Then, as in previous experiments, all participants answered open-ended general knowledge questions related to the information they read in the tweets. We examined the rate at which participants reproduced false

information and produced correct answers to those questions, and whether these patterns differed based on the task participants engaged in during exposures to the tweets.

Several possible patterns could emerge as a function of the different judgments. One possibility is that asking people to consider whether they would like the tweets may prompt self-monitoring as compared to being tasked with making a non-evaluative, interest judgment. Previous work has suggested that people are concerned about the accuracy of information that they engage with online to avoid negative perceptions associated with interacting with false information (Altay et al., 2020; Waruwu et al., 2020). Pennycook, Epstein, et al. (2021) found that over 80% of surveyed participants said that they thought that the accuracy of posts was “very” or “extremely” important when making decisions about what to share on social media, which outranked every other content dimension (i.e., political alignment, interest, funniness, surprise). It is therefore possible that being in a social media mindset, as prompted via “like” judgments, could direct people’s attention to the accuracy of the presented information, as those decisions can be seen as more reflective of online reputations and experiences. Thus, the pattern of responses following like decisions should be quite similar to those obtained following accuracy judgments. This would result in fewer inaccurate reproductions and more correct answers in the social media mindset and accuracy mindset conditions than in the interest mindset condition.

Another possibility is that evaluating tweets for accuracy at exposure could reduce incorrect lure reproductions and increase correct answer production as compared to the other two conditions. People do not seem to recognize the need to engage in evaluation, or opt to evaluate information, without sufficient prompting (Salovich & Rapp, 2021). Research has in fact consistently shown that people often fail to adopt an accuracy focus under normal reading

conditions (e.g., see Rapp & Braasch, 2014; Rapp, 2016 for reviews). Even giving participants a goal of avoiding falsehoods (Andrews-Todd et al., 2021), or informing them about the possibility of encountering inaccuracies (Jalbert et al., 2018; Marsh & Fazio, 2006; Salovich & Rapp, 2022), can be insufficient for motivating evaluation. Others have found that people rarely take the accuracy of news headlines into account when making sharing decisions, despite self-reported importance of considering accuracy when interacting with content online (Pennycook, Epstein, et al., 2021). This suggests that an accuracy focus is unlikely to spontaneously occur even when participants are asked to make more realistic and relevant social media judgments. Interest mindset and social media mindset conditions may therefore lead to similar rates of incorrect and correct responses, with reduced susceptibility to inaccuracies largely constrained to the accuracy mindset condition when evaluation is explicitly prompted.

It is also of course possible that no differences could emerge across mindset conditions. We suspected this was unlikely given consistent differences have emerged in previous work between participants prompted to consider the interest versus accuracy of potentially false information (e.g., Brashier et al., 2020; Calvillo & Smelter, 2021; Salovich, Kirsch, et al., 2022). Thus, a crucial addition here involves what happens in the “like” condition, and whether it might resemble the interest or accuracy conditions.

Method

The pre-registration for this experiment can be found at https://aspredicted.org/V4Y_18K.

Participants

Sample size was calculated based on a SESOI (i.e., smallest effect of interest) simulation using the mixedpower package in R (Kumle et al., 2021). In the simulation we included SESOIs

in the form of beta coefficients that were 30% smaller than those returned by the model used to assess the effects of judgment type in Salovich, Kirsch, et al., (2022). This was due to observed differences in effect size between Experiment 6 (the effect of reading true, false, and filler tweets) and Salovich, Kirsch, et al.'s (2022) Experiment 1 (the effect of reading true, false, and filler statements). The simulations suggest $\sim n = 140$ participants are required per between-subject judgment condition to reach .80 power to detect the statement difficulty x judgment type interactions of interest. As such, we recruited four hundred and twenty ($n = 140$ per three between-subject conditions totaling 30,240 observations) Amazon Mechanical Turk workers to participate in the study (192 female; M age = 39.47 years). We used the same exclusion criteria as in Experiment 6 and continued recruitment until the pre-registered target sample size was reached. All participants were paid equal to or above the U.S. minimum wage at the time of data collection.

Participants on average reported using 4.15 different social media apps with YouTube (87%) and Facebook (74%) being most popular platforms. Over two-thirds (70%) of participants reported using Twitter. Most (64%) respondents reported using social media between 10-30 minutes per day (16%), 31-60 minutes per day (21%), or 1-2 hours per day (27%).

Design

The experiment followed a 3 (mindset condition: interest, accuracy, social media) x 2 (tweet validity: true or false) x 2 (tweet difficulty: easy or hard) mixed design. Mindset condition was manipulated between groups, while tweet validity and difficulty were manipulated within-subjects. Participants were asked knowledge questions related to all 72 tweets, allowing us to measure rates of incorrect lure and correct answer responses. In addition to the general knowledge questionnaire, we also measured the time it took participants to complete each rating

at exposure and each judgment at test, and as well collected open-response self-reports of how participants decided to rate tweets on interest, accuracy, or whether they would like them.

Materials

The same materials were used as in Experiment 6 except we omitted filler items, presenting participants only with either the true or false version of each of the 72 tweets. All variations were orthogonal, in that participants viewed an equal number of easy true and false tweets, and an equal number of hard true and false tweets.

Procedure

Participants engaged in a similar procedure as in Experiment 6, with an exposure phase to true and false (but not filler) tweets that was followed by a test phase of related open-answer questions. Participants received the same initial instructions as in Experiment 6, and then were assigned to one of three between-subject judgment conditions: During the exposure phase, participants in the *interest mindset* condition rated each tweet on interest from 1 (uninteresting) to 5 (interesting); participants in the *accuracy mindset* condition rated each tweet on how accurate it was from 1 (false) to 5 (true); participants in the *social media mindset* condition rated each tweet on how likely they would be to ‘like it’ if they saw it on Twitter from 1 (unlikely) to 5 (likely). Each tweet was presented one-at-a-time in a unique random order for each participant. Time taken to complete each judgment was captured in milliseconds, beginning at the time that the tweet was displayed and concluding when participants submitted their response.

Then, in the test phase, participants answered each general knowledge question one-at-a-time, presented in a unique random order for each participant. They again were asked to answer to the best of their knowledge without referencing external resources and were allowed to leave

answers blank. Time taken to submit a response to each of the general knowledge questions was again captured in milliseconds from the time of display to time of submission.

Finally, we collected demographics and information on participants' social media activity and platform use with the same questions as in Experiment 6. At this time, we also asked participants to explain in a few sentences what considerations went into their interest, accuracy, or like judgments, depending on condition. Participants submitted their responses by typing in an open text box.

Results

Questionnaire Coding

All responses ($N = 30,240$) were coded using the same procedure and criteria as in Experiment 6. Inter-rater reliability for dual-coded responses was $\kappa = .90$ with all disagreements resolved through discussion.

Models

Open-response data were analyzed using GLMM models as in Experiment 6, with separate examinations of incorrect lure productions and correct productions. Other exploratory analyses used GLMMs as well as linear mixed effect models (LMM) using the R packages lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2014). All model specifications and outputs for Experiment 7 are publicly available at <https://rpubs.com/nsalovich/twitter-exp2>.

Questionnaire Responses

Incorrect Lure Productions. Incorrect lure reproductions are summarized in Table 8 and Figure 14. First, we ran a GLMM predicting incorrect lure responses (coded as 1) with tweet validity (contrast coded; true = -.5, false = .5) and tweet difficulty (contrast coded; easy = -.5 and hard = .5) as fixed effects, participant and item as random intercepts, and tweet validity and

difficulty as random slopes. As expected, participants produced more incorrect lures following exposure to tweets containing those inaccuracies ($M = .11$, $SD = .11$) than tweets containing true information ($M = .02$, $SD = .03$), $b = 1.86$, $z = 22.54$, $p < .001$. While participants numerically produced more incorrect lures for hard ($M = .08$, $SD = .10$) versus easy items ($M = .05$, $SD = .09$), the main effect of tweet difficulty did not reach significance, nor was there a significant difficulty x validity interaction, $ps > .05$. People were influenced by exposure to both well-known and lesser-known inaccuracies.

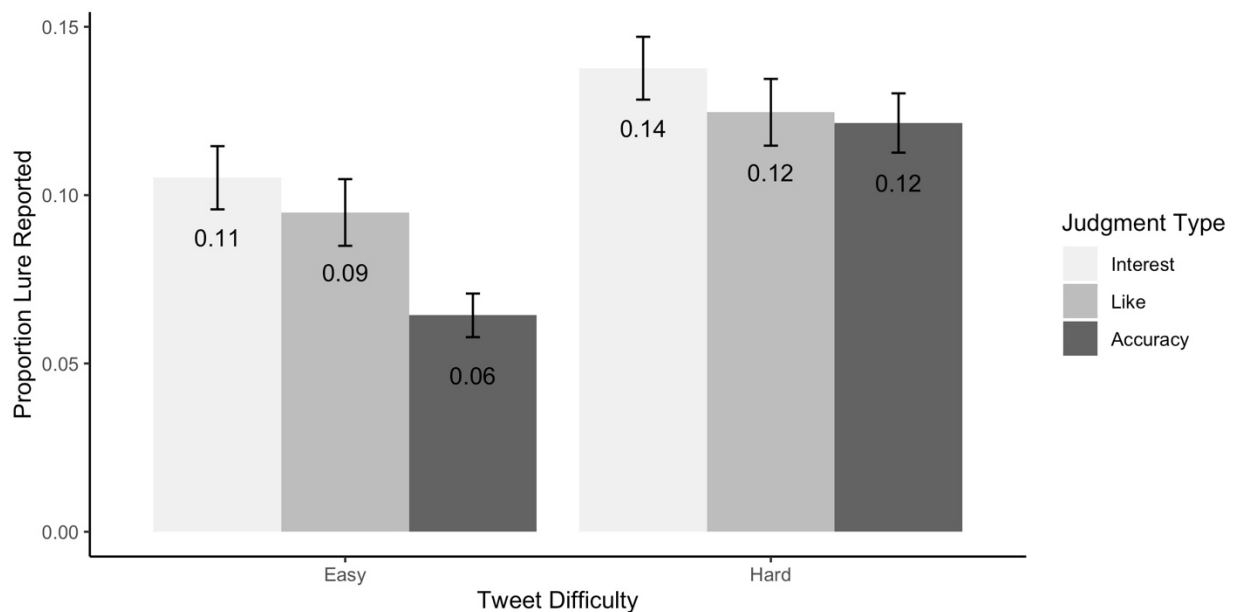
We then investigated how mindset condition affected reproductions of inaccuracies. We ran a GLMM predicting incorrect lure reproduction following exposure to tweets containing false information with tweet difficulty (contrast coded; easy = -.5 and hard = .5) and mindset condition as fixed effects, participants and item as random intercepts, and tweet difficulty as a random slope. We used a pair of orthogonal contrasts to test predicted differences between mindset condition reflected in our pre-registered hypotheses. The first contrast tested whether the accuracy mindset condition (coded as -2) produced fewer incorrect lures than both interest and social media mindset conditions (coded as 1, respectively), and the second contrast tested whether any differences emerged between interest (coded as 1) and social media mindset conditions (coded as -1).

As predicted, participants in the accuracy mindset condition overall reproduced fewer incorrect lures ($M = .09$, $SD = .10$) as compared to participants in the interest mindset ($M = .12$, $SD = .11$) and social media mindset conditions ($M = .11$, $SD = .12$), $b = .11$, $z = 2.86$, $p = .004$. This effect was qualified by a significant interaction with tweet difficulty, $b = -.15$, $z = -3.02$, $p = .003$. Simple effects revealed that this difference emerged for easy, well-known information (accuracy: $M = .06$, $SD = .08$; interest: $M = .11$, $SD = .11$; social media: $M = .10$, $SD = .12$), $z = -$

3.85, $p < .001$, but not for hard, lesser-known information (accuracy: $M = .12$, $SD = .10$; interest: $M = .14$, $SD = .11$; social media: $M = .13$, $SD = .12$), $z = -.80$, $p = .43$. No differences emerged between incorrect lures reproduced between interest and like judgment conditions, nor did any other interactions emerge, $ps > .05$. Accuracy mindsets reduced reproductions of false information from tweets, but only for information that people likely possessed accurate prior knowledge about (see Figure 14).

Figure 14

Error Rates Following False Information in Experiment 7



Note: Error rates following exposure to false information split by tweet difficulty (easy or hard) and mindset condition (interest, social media, accuracy). Error bars represent standard error.

Correct Answer Productions. Correct response rates are summarized in in Table 8 and Figure 15. To analyze correct answers, we first ran a GLMM predicting correct responses (coded as 1) with tweet validity (contrast coded; true = .5, false = -.5) and tweet difficulty (contrast

coded; easy = .5 and hard = -.5) as fixed effects, participant and item as random intercepts, and tweet validity and difficulty as random slopes. As expected, participants were more likely to produce correct responses to easy ($M = .82, SD = .20$) than hard questions ($M = .34, SD = .28$), $b = 3.16, z = 17.55, p < .001$. They also produced more correct responses after exposure to tweets containing true ($M = .64, SD = .33$) as compared to false information ($M = .52, SD = .33$), $b = 1.04, z = 21.65, p < .001$. There was also a significant statement validity x statement difficulty interaction, $b = .38, z = 4.95, p < .001$. As in previous work, analysis of simple effects revealed that participants were more likely to produce correct answers after exposure to true than to false tweets when the statement contents were familiar or “easy”, $z = 18.69, p < .001$, versus unfamiliar or “hard”, $z = 15.13, p < .001$.

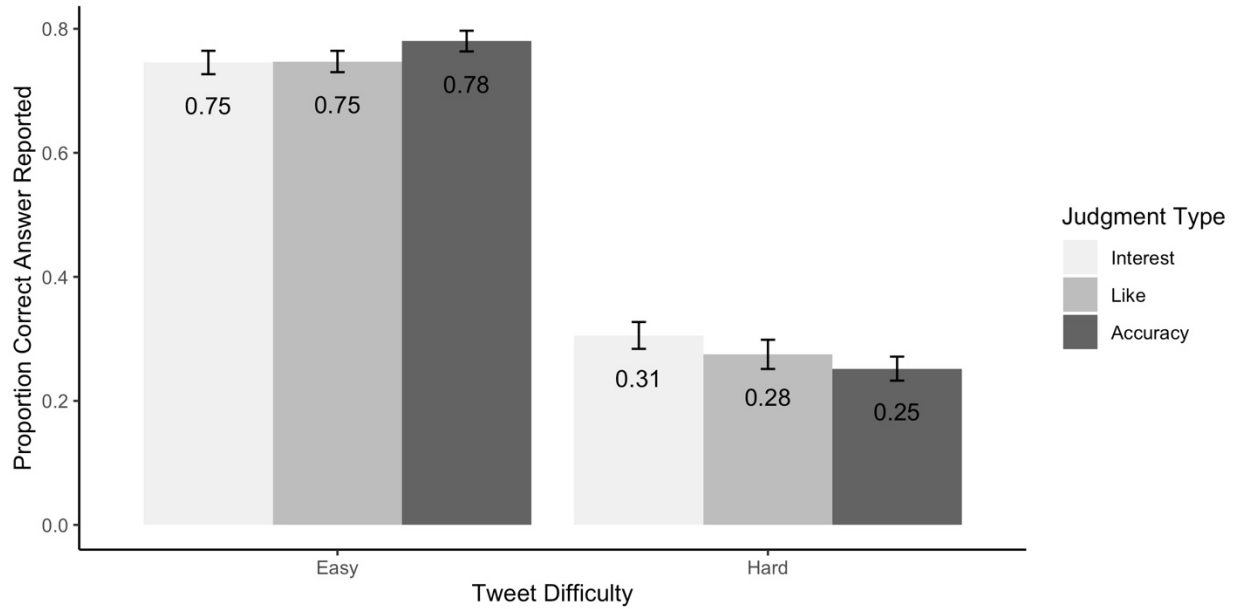
We next investigated how mindset at exposure affects correct responses. We ran a GLMM predicting correct answers produced after exposure to tweets containing false information with tweet difficulty (contrast coded; easy = .5 and hard = -.5) and mindset condition as fixed effects, participants and item as random intercepts, and tweet difficulty as a random slope. Again, we used a pair of orthogonal contrasts to test predicted differences between mindset condition reflected in our pre-registered hypotheses. The first contrast tested whether the accuracy mindset condition (coded as 2) produced more correct answers than both the interest and social media mindset conditions (coded as -1, respectively), and the second contrast tested whether any differences emerged between the interest (coded as -1) and social media mindset conditions (coded as 1).

Although no overall differences in correct answers emerged between the accuracy mindset ($M = .48, SD = .50$) versus interest ($M = .48, SD = .50$) and social media mindset conditions ($M = .48, SD = .50$), $p > .05$, there was a significant interaction between this contrast

and tweet difficulty, $b = .48$, $z = 3.01$, $p = .003$. To interpret this interaction, we conducted simple effects test considering the differences in mindset conditions when tweets were easy or hard, respectively. Accuracy mindset numerically increased correct answers ($M = .78$, $SD = .20$) relative to interest ($M = .75$, $SD = .22$) and social media mindset conditions ($M = .75$, $SD = .20$) when the information was easy, $z = 1.31$, $p = .19$, but the opposite trend emerged for hard statements (accuracy: $M = .25$, $SD = .23$, interest: $M = .31$, $SD = .26$, social media mindset: $M = .28$, $SD = .28$), $z = -1.56$, $p = .12$; $z = -1.56$, $p = .12$. While neither simple effect on its own reached significance, the differential patterns of effects of accuracy judgments when the tweets were easy versus hard resulted in the significant interaction (see Figure 15). As in analyses of incorrect lures, no differences emerged between interest and social media mindset conditions, $ps > .05$. Evaluating false tweets increased correct responses when the information was likely to contradict participants' accurate prior knowledge, but otherwise led to reduced productions of correct answers (see Figure 15).

Figure 15

Correct Response Rates Following False Information in Experiment 7



Note: Correct response rates following exposure to false information split by tweet difficulty (easy or hard) and mindset condition (interest, social media, accuracy). Error bars represent standard error.

Table 8

Mean rates of incorrect lure and correct response production in Experiment 7

Mindset Condition	Tweet Difficulty	Lure	Correct
<i>Interest</i>			
True	Easy	.02 (.04)	.87 (.18)
	Hard	.03 (.03)	.42 (.29)
False	Easy	.11 (.11)	.75 (.22)
	Hard	.14 (.11)	.31 (.26)
<i>Social Media</i>			
True	Easy	.01 (.03)	.88 (.15)
	Hard	.03 (.03)	.40 (.29)
False	Easy	.09 (.12)	.75 (.20)
	Hard	.12 (.12)	.28 (.28)
<i>Accuracy</i>			

True	Easy	.01 (.03)	.88 (.20)
	Hard	.03 (.03)	.40 (.27)
False	Easy	.06 (.08)	.78 (.20)
	Hard	.12 (.10)	.25 (.23)

Note: Numbers in parentheses are standard deviations.

Response Times

To examine whether there were differences in response times between mindset conditions, we ran two LMMs predicting judgment response time at exposure and answer response time at test by mindset condition, with participant and item as random intercepts, and mindset condition as a random slope. The accuracy mindset condition was set as the referent condition due to focus on how evaluative, accuracy judgments affected response time as compared to non-evaluative, interest and like judgments, consistent with our analyses of questionnaire responses.

During the exposure phase, participants took significantly longer to judge tweets for accuracy ($M = 11.58$, $SD = 22.76$) as compared to interest ($M = 9.07$, $SD = 20.59$), $b = -2.51$, $t = -3.61$, $p < .001$, or whether they would like the tweets ($M = 9.90$, $SD = 26.52$), $b = -1.69$, $t = -2.34$, $p = .02$. There were, however, no significant differences in time taken to answer the general knowledge questions at test between the accuracy mindset ($M = 9.06$, $SD = 19.14$), interest mindset ($M = 10.27$, $SD = 19.52$), or social media mindset conditions ($M = 9.54$, $SD = 15.97$), $ps > .05$.

Relationship Between Exposure Judgment and Test Answers

We next analyzed whether the accuracy, interest, or like judgments made at exposure were predictive of whether participants would reproduce that information on the general knowledge questionnaire. We ran three separate GLMMs predicting whether participants reproduced answers as a function of their rating of the content at exposure in each respective

mindset condition. In all models, participant and item were included as random intercepts, and judgment rating was included as a random slope. This accounts for the possible variability in how each participant rated the tweets (e.g., overall found tweets more or less interesting), as well as how each tweet was rated across participants (e.g., overall one tweet was found more or less interesting).

People who rated tweets as more accurate were more likely to reproduce those answers at test, $b = .23$, $z = 2.44$, $p = .01$. However, neither interest ratings nor like ratings at exposure were predictive of later reproductions, $ps > .05$.

Accuracy as a Self-Reported Strategy Underlying Judgments

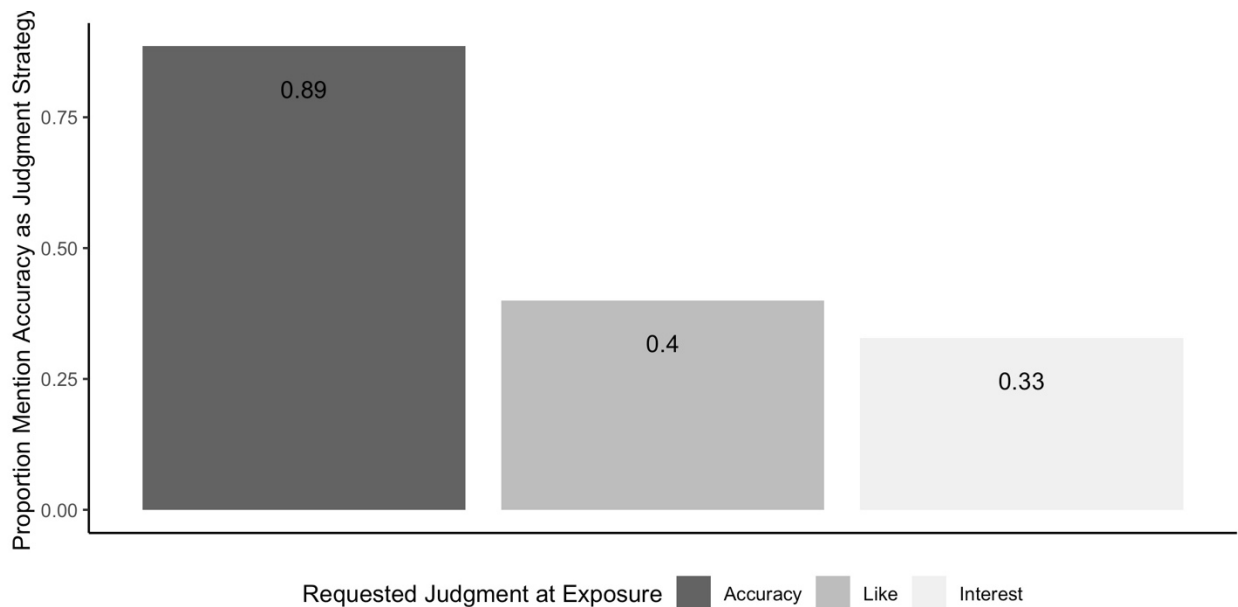
To examine whether participants considered accuracy as a strategy when reading the true and false tweets and making their judgments, we completed a content analysis of their free responses. We examined when participants explicitly reported that accuracy was a consideration for their initial judgments during the exposure phase of the study. We note that these free responses were collected at the conclusion of the experiment after participants had completed both the exposure and test phases. Two raters independently coded a quarter of the responses in the data set, with the remaining coded by one rater only. Both raters were blinded to judgment condition. Inter-rater reliability for dual-coded responses was reliably high ($\kappa = .98$) with all disagreements resolved through discussion. While we only present analyses for whether or not accuracy was mentioned in the response, all open-responses are available on OSF for future content analyses of these qualitative reports.

The proportion of subjects who reported considering accuracy in their judgments differed by mindset condition, $X^2(2,420) = 103.53$, $p < .001$. Post-hoc pairwise tests using the R package “fifer” (Fife, 2014) indicated that participants in the accuracy mindset condition reported

considering the accuracy of the tweet's contents when making their judgments 89% of the time, which is significantly more often than in the interest mindset condition (33%) and in the social media mindset condition (40%), $ps < .001$ (see Figure 16). There were no differences in accuracy mentioned as a strategy between the interest and social media mindset conditions, $p = .26$.

Figure 16

Self-Reported Consideration of Accuracy in Experiment 7



Note: Proportion of participants who reported considering the accuracy of information in the interest, accuracy, and social media mindset conditions.

We next considered whether mentions of accuracy as a self-reported strategy predicted the degree to which participants reproduced incorrect lures and produced correct answers following exposures to false information (see Table 9 for summary). First, we ran a GLMM predicting incorrect lures reproduced at test by statement difficulty (contrast coded; easy = -.5 and hard = .5) and whether participants' free response mentioned accuracy considerations (contrast coded; mentioned accuracy = -.5, did not mention accuracy = .5). As in prior analyses,

both participants and items were entered as fixed effects and difficulty as a random slope.

Participants who had self-reported accuracy as a strategy in making their initial judgments ($M = .10$, $SD = .10$) were less likely to reproduce incorrect lures than were participants who did not mention accuracy in their response ($M = .12$, $SD = .12$), $b = .31$, $z = 2.87$, $p = .004$. A significant mindset condition x statement difficulty interaction emerged, $b = -.95$, $z = -7.27$, $p < .001$.

Simple contrasts revealed that, as compared to participants who did not mention accuracy, participants who reported considering accuracy produced fewer incorrect lures to answer easy questions, $z = 6.08$, $p < .001$, but with no differences for hard items, $z = -1.36$, $p = .17$. Both main, $b = 1.14$, $z = 6.67$, $p < .001$, and interaction effects, $b = -1.09$, $z = -6.49$, $p < .001$, remained significant even when restricting analysis to participants in the interest and social media mindset conditions who spontaneously identified accuracy as a strategy.

To examine effects on correct responses, we ran a GLMM predicting correct answers produced at test by statement difficulty (contrast coded; easy = .5 and hard = -.5) and whether participants' free responses mentioned accuracy considerations (contrast coded; mentioned accuracy = .5, did not mention accuracy = -.5). Again, participants and items were entered as fixed effects and difficulty as a random slope. Participants who self-reported accuracy as a strategy for their initial judgments ($M = .57$, $SD = .33$) were more likely to respond with correct answers than were participants who did not mention accuracy in their response ($M = .46$, $SD = .33$), $b = .81$, $z = 6.04$, $p < .001$. A significant mindset condition x statement difficulty interaction $b = .40$, $z = 2.59$, $p = .01$ indicated this difference was greater when the information was easy, $z = 7.99$, $p < .001$, versus hard, $z = .61$, $p < .001$. While the main effect of mentioning accuracy was significant when constraining analyses to interest and social media mindset conditions, $b = 1.14$, $z = 6.67$, $p < .001$, the interaction with statement difficulty was not, $p > .05$.

Table 9

Mean Rates of Incorrect Lure Reproduction and Correct Response Production Based on Self-Reported Considerations of Accuracy Experiment 7

Mindset Condition and Tweet Difficulty	Mentions Accuracy	Lure	Correct
<i>Interest</i>			
Easy	No	.12 (.12)	.70 (.24)
	Yes	.07 (.07)	.85 (.15)
Hard	No	.13 (.10)	.27 (.26)
	Yes	.16 (.13)	.38 (.23)
<i>Social Media</i>			
Easy	No	.13 (.13)	.67 (.21)
	Yes	.05 (.07)	.87 (.12)
Hard	No	.12 (.13)	.20 (.25)
	Yes	.13 (.10)	.38 (.29)
<i>Accuracy</i>			
Easy	No	.07 (.09)	.60 (.27)
	Yes	.06 (.08)	.80 (.17)
Hard	No	.07 (.08)	.24 (.24)
	Yes	.13 (.11)	.25 (.23)

Note: All means following exposures to tweets containing false information only. Numbers in parentheses are standard deviations.

Discussion

As in Experiment 6, participants were more likely to produce incorrect lures and less likely to respond with correct answers when previously exposed to tweets containing false as compared to true information. Evaluating the accuracy of the tweets at exposure significantly attenuated the problematic consequences of exposures to the inaccuracies: Participants asked to make accuracy decisions were less likely to reproduce false ideas and more likely to produce correct responses than were participants asked to make interest or liking decisions. No differences emerged, however, between either non-evaluative mindset condition. As in previous

projects, adopting an accuracy focus was most beneficial when the presented ideas were easy or familiar, as participants were able to effectively determine these claims to be false based on their accurate prior knowledge (e.g., Salovich, Kirsch, et al., 2022).

Exploratory analyses corroborated the unique importance of contemplating accuracy for reducing the effects of reading inaccurate information. Participants prompted with accuracy judgments took significantly more time to evaluate the tweets for validity than participants who were tasked with considering their interest in the tweets or participants who were tasked with considering whether they would “like” the tweets if they encountered them on social media. The longer judgment times might be suggestive of participants’ detecting discrepancies in the inaccuracies (Rapp, 2008; Richter, 2015; Singer, 2013, 2019; Weil & Mudrik, 2020), and/or retrieval of correct prior knowledge used to evaluate the validity of presented ideas (Rapp, Hinze, et al., 2014; Salovich, Kirsch, et al., 2022), as compared to response latencies in for the non-evaluative judgments. Similar latency differences have been offered as evidence for different in-the-moment strategies people leverage when readers process content given different reading goals and tasks (Rapp & Mensink, 2011). Furthermore, participants’ accuracy judgments at exposure were predictive of their subsequent reproductions of that information; tweets rated as more or less interesting, or more or less likely to be “liked,” were not indicative of subsequent use of that information. This suggests that the accuracy mindset encouraged by evaluating the statements prompted a unique focus on the validity of information, resulting in downstream consequences that differed from the interest and social media mindset conditions.

Participants’ self-reported accuracy considerations further inform and align with these findings. Perhaps unsurprisingly, the vast majority of participants tasked with accuracy judgments mentioned considering the validity of the tweet content and its alignment with their

prior knowledge, and did so significantly more often than did participants tasked with making interest or “like” judgments. Yet regardless of assigned judgment condition, participants who mentioned accuracy as a factor reproduced significantly fewer false ideas and produced more correct answers than did participants who did not mention accuracy. Again, these benefits were obtained specifically when information was familiar to participants. This further underscores the clear and consistent benefits associated with adopting an accuracy mindset for protecting against the influence of false information.

General Discussion

A growing body of work has been dedicated to determining when and why people rely on false or misleading information (Brashier & Marsh, 2020; Britt et al., 2019; Lewandowsky et al., 2017; Pennycook & Rand, 2021; Rapp & Braasch, 2014; Rapp, 2016; Salovich & Rapp, 2018), including in cases when they should “know better” based on their accurate prior knowledge (Fazio et al. 2015; Fazio et al., 2019; Salovich, Kirsch, et al., 2022). Many of these investigations have relied on decontextualized single sentences, either in the form of true or false declarative facts (e.g., [Jupiter/Saturn] is the largest planet in the solar system), or accurate or inaccurate assertions about real-world occurrences (e.g., toothbrushing [prevents/causes] gum disease). While these ideas are not unlike those that would be encountered during daily discourse experiences, there are arguably few situations in which people might come across these kinds of claims as isolated true and false statements. It is therefore important to determine whether the effects observed in those studies also emerge with other kinds of materials that regularly occur in real-world contexts. Some researchers have begun to tackle this issue by assessing the influence of inaccurate claims embedded in fictional narratives (e.g., Gerrig & Prentice, 1991; Marsh & Fazio, 2006; Marsh et al., 2003; Rapp, 2008; Salovich et al., 2021; Salovich, Imundo, et al.,

2022), which nicely aligns with the tendency for authors to include exaggerated or inaccurate ideas in their storylines to entertain readers. Others have focused on characterizing people's decisions with respect to dissemination of true and fake news headlines (e.g., Brashier et al., 2021; Calvillo & Smelter, 2021; Pennycook et al., 2020; Pennycook, Epstein, et al., 2021), which also represent single statements, but in practice provide more contextual cues (e.g., photos, webpage source) and link to more expanded write-ups (e.g., full-length articles). Few projects have been dedicated to understanding the effects of reading false claims and ideas on social media posts despite these experiences being routine, and despite their offering a host of falsehoods or half-truths given their disintermediated origins (Britt et al., 2019; Lazer et al., 2018; Lewandowsky et al., 2017; Pennycook & Rand, 2021; Shoemaker & Vos, 2009). To our knowledge, this is one of the first projects dedicated to assessing whether, when, and how people encode and reproduce ideas presented through social media that run counter to their existing accurate understandings of the world.

The purpose of Chapter 3 was to extend the findings obtained in previous projects, including the experiments summarized in Chapters 1 and 2, to a social media context. In Experiment 6, we examined whether and how false information presented in tweets influences people's responses to general knowledge questions. As in previous work (e.g., Donovan & Rapp, 2020; Eslick et al., 2011; Fazio Dolan, et al., 2015; Marsh & Fazio, 2006; Marsh et al., 2003; Salovich, Kirsch, et al., 2022), but this time using stimuli modeled after tweets, exposures to false information increased the rate at which inaccuracies were reproduced by participants to answer general knowledge questions relative to reading true information or no relevant information. These problematic effects emerged for both well-known and lesser-known inaccuracies. Exposures to false content appearing in tweets also reduced participants' correct

answers relative to viewing true tweets or no information, particularly when the information was familiar. This is consistent with the idea that reading inaccuracies can problematically cause people to doubt and even discount their accurate prior knowledge (Rapp & Salovich, 2018; Salovich, Kirsch, et al., 2022). Together these effects crucially demonstrate how even brief exposures to inaccurate information on social media can impact what people believe or at least report to be true about the world.

In Experiment 7 we expanded on these findings by examining how tasks and mindsets that people might adopt while interacting with information on social media might affect their reliance on that content. In particular, we tested whether adopting an accuracy mindset would protect people from the influence of false ideas presented within tweets. Participants judged realistic tweets for accuracy, interest, or whether they would like the tweet—the latter of which was designed to mirror the type of contexts and decisions that participants routinely engage in when consuming information on social media. Participants who judged each tweet at exposure for accuracy reproduced significantly fewer incorrect ideas from those tweets than did participants tasked with making either non-evaluative interest or like judgments about the same information. Accuracy judgments also increased participants' use of correct prior knowledge in lieu of the presented inaccuracies when those understandings were likely available. This is consistent with prior work demonstrating that accuracy judgments prompt evaluative mindsets that can protect against the problematic effects of exposure to false information (Brashier et al., 2020; Calvillo & Smelter, 2021; Pennycook et al., 2020; Pennycook, Epstein, et al., 2021; Rapp Hinze, et al., 2014; Salovich, Kirsch, et al., 2022). In the current set of studies, we demonstrated that an accuracy mindset can reduce people's actual reproduction of false information, including ideas communicated via realistic tweets, encouraging them to instead rely on accurate prior

knowledge. This offers crucial support for the utility of interventions that prompt an accuracy focus in reducing reliance on falsehoods in real-world contexts and experiences (Pennycook & Rand, 2022).

Besides helping to establish the generalizability of prior effects, the current project also interrogated the downstream effects of encountering inaccurate information under *non-evaluative* tasks and mindsets. Previous studies that have assessed the utility of prompting an accuracy focus have traditionally only contrasted its effects with conditions in which participants are prompted to rate their interest in the presented information (e.g., Brashier et al., 2020; Calvillo & Smelter, 2021; Salovich, Kirsch, et al., 2022). The assumption is that such tasks act as a neutral control to encourage comprehension without necessarily prompting accuracy evaluations. But instantiating an interest focus may actually encourage shallow processing strategies, potentially dissuading considerations that may have otherwise emerged under more naturalistic experiences, including potential evaluation (Hawkins & Hoch, 1992). As such, the benefits associated with an accuracy focus may actually be unduly overestimated when contrasted with non-evaluative interest judgments. In Experiment 7, we begin to address this concern by contrasting performance following accuracy and interest judgments with a third judgment that routinely occurs on social media: whether they would “like” each tweet if seen on their twitter timelines. We leveraged this social media judgment to realistically reflect choices people may make about online content, which might potentially motivate more evaluative considerations than arbitrary interest judgments. To test this, we examined whether in-the-moment considerations of accuracy and/or downstream reliance on false ideas produced by participants prompted with social media judgments were more similar to those of participants prompted with non-evaluative interest judgments, or those who were prompted with evaluative accuracy judgments.

Participants tasked with making interest or liking judgments showed similar rates of inaccurate reproductions and similar rates of correct answers at test. The latter two non-evaluative mindset conditions were associated with notably greater reliance on presented inaccuracies than was the accuracy mindset condition, which appeared to successfully attenuate traditionally reported problematic effects. While this does not necessarily mean the exact same processes and considerations underlie both interest and “like” judgments, in-the-moment and qualitative measures support the idea that similar strategies may have been operating in these two conditions. First, judgment latencies did not differ between the interest and social media mindset conditions, and both took significantly less time to complete than did accuracy judgments. The latter two groups also reported similarly low rates of spontaneous accuracy considerations for making their initial judgments about the information. This was observed despite prior projects highlighting that participants frequently self-disclose accuracy as an important factor underlying their interactions with online content (Altay et al., 2020; Pennycook, Epstein, et al., 2021). Further, only participants’ initial accuracy judgments were predictive of their later use of the presented information, whereas people’s interest or propensity to like the tweet were unrelated to subsequent responses. These measures corroborate that evaluative mindsets underlying accuracy decisions can protect against the influence of false information on social media, and suggest that non-evaluative tasks and mindsets, including the interest judgments integral as a control in prior investigations, result in similar, increased susceptibility to falsehoods.

Future work should continue to characterize the various kinds of judgments, tasks, and mindsets that can encourage or dissuade reliance on inaccuracies information presented in a social media context. While the results of Experiment 7 indicate that non-evaluative considerations may result in susceptibility to false ideas, previous work has identified a multitude

of underlying factors that people prioritize when deciding what to like, share, and comment while online, including but not limited to humor, political alignment, and surprise (e.g., Pennycook, Epstein, et al., 2021). Additionally, differences can emerge in people's spontaneous considerations of accuracy in deciding what to like versus what to share with others, as the latter incurs greater social and reputational concerns, and therefore may be more likely to encourage careful evaluation (e.g., Altay et al. 2020; Brashier & Schacter, 2020; Effron, 2018; Waruwu et al., 2021). Researchers should also investigate the possibility that social media in general may dissuade considerations of accuracy. As a result of competing distractions and/or mindless scrolling through a wide array of content, social media users may be primed to adopt goals and mindsets aligned with entertainment experiences over more epistemic concerns. This could encourage individuals to "let their guard down" when consuming information as compared to in other more expository environments (Corneille et al., 2020; Salovich & Rapp, 2021; Salovich & Rapp, 2022). Although people may prioritize a variety of factors and considerations when engaging with information online, recent work summarized in Chapters 1 and 2 suggests that evaluative benefits can emerge and persist over and above existing non-evaluative judgments and goals (e.g., Salovich, Kirsch, et al., 2022; Salovich & Rapp, 2022). This acts as further support for the viability of accuracy nudges in reducing the belief and spread of false information in contexts like social media (Pennycook & Rand, 2022). More work is needed to characterize the processes and products of our social media engagements and how they may interact with these interventions.

Understanding and addressing the negative ramifications of misinformation and "fake news" continues to be a priority for researchers, political and policy leaders, and the general public alike. In two studies we contribute to these timely examinations by demonstrating that

exposure to even blatant inaccuracies in tweets can problematically increase reproductions of that information and decrease reliance on accurate prior knowledge. This aligns with concerns about the spread and influence of inaccurate information at the forefront of both theoretical and everyday discussions. We also find promising evidence that the problematic influence of inaccurate information can be reduced by urging participants to adopt an evaluative mindset and contemplate the accuracy of potentially false or misleading ideas. This does not appear to occur spontaneously during interactions with social media content, which makes sense given the diverse reasons people have for engaging with social media. Follow-up projects should examine the extent to which these effects replicate across other popular and influential social media platforms (e.g., Facebook, Instagram, TikTok), as well as continue to investigate and test the efficacy of interventions using actual social media data (e.g., Pasquetto et al., 2020; Pennycook, Epstein, et al., 2021). With crises like the COVID-19 pandemic exacerbated by the circulation of false information (Loomba et al., 2021), it is pressing to not only understand the cognitive processes that contribute to people's reliance on false or misleading claims, but also when and how they can be leveraged in the design of interventions to support a more informed society.

Conclusion

The purpose of this dissertation was to better understand how the tasks and mindsets with which people approach information affect the likelihood they are influenced by false claims and ideas. In Chapter 1, participants rated true and false statements (e.g., Jupiter/Saturn is the largest planet in the solar system) for accuracy (i.e., an evaluative judgment) or interest (i.e., a non-evaluative judgment). In Experiment 1, participants who evaluated statements for accuracy were less likely to reproduce ideas from the false statements (“Saturn”) and more likely to rely on correct prior knowledge (“Jupiter”) to answer post-reading questions than were participants who made non-evaluative interest judgments. In Experiments 2 and 3, the same benefits were obtained even when participants were not consistently prompted to evaluate the statements. Results from these three experiments offer insight into when and how evaluation can encourage participants to rely on correct prior knowledge over presented inaccuracies, as well as what is required to establish and maintain such an evaluative mindset.

In Chapter 2, we examined whether people can be encouraged to develop and maintain evaluative mindsets without explicit instruction to do so. We tested whether and how confronting people about their potential susceptibility to the influence of false information might motivate evaluation and reduce these problematic effects. Participants made non-evaluative interest ratings about true and false statements, and then answered related general knowledge questions. In Experiment 4, participants who received positive or negative performance feedback about their susceptibility to inaccurate information reproduced fewer incorrect ideas and produced more correct answers than did participants who did not receive feedback. In Experiment 5, analogous benefits emerged when participants were simply informed that their use of false information was being monitored. These results highlight the importance of people’s thoughts

and beliefs about their own resistance to inaccuracies and with respect to their actual performance, suggesting critical mechanisms underlying the effects associated with exposures to falsehoods. The findings inform practical approaches for supporting accurate knowledge acquisition in a variety of settings, without necessarily relying on explicit instruction or direction to engage in evaluation.

Chapter 3 extends the findings obtained in previous projects, exploring how both evaluative and non-evaluative tasks and mindsets might differentially affect the influence of inaccuracies as presented in a social media context. We modified the statements used in the experiments from Chapters 1 and 2 into realistic tweets containing either correct or incorrect general knowledge information. Using these newly-developed stimuli, we conducted two experiments examining how the tasks and mindsets with which people approach information on social media can affect the degree to which they are susceptible to false information. First, in Experiment 6, we found that exposures to false information in tweets increased the rate at which those inaccuracies were reproduced by participants to answer general knowledge questions relative to viewing true information or no relevant information. Exposures to false tweets also problematically reduced correct answer productions relative to viewing true tweets or no information. In Experiment 7, we demonstrated that participants who judged tweets on accuracy at exposure (associated with adopting an accuracy mindset) were less likely to reproduce inaccuracies from tweets and more likely to answer with their accurate prior knowledge than those who rated how interesting the tweets were (associated with an interest mindset), and how likely they would be to “like” the tweet on social media (which we label a social media mindset). Exploratory analyses of response times, the relationship between ratings of interest, accuracy, and “liking” a tweet and participants’ reproductions of presented information, and participants’

qualitative testimonies on their accuracy contemplations corroborated the unique importance of an accuracy mindset for reducing the effects of reading inaccurate information. Together these studies further our understanding of the circumstances that encourage and dissuade evaluation in online contexts. This offers insight into the processes that underlie people's susceptibility to false information, and informs potential real-world applications for reducing the dissemination of misinformation on social media.

The experiments presented in this dissertation contribute to contemporary accounts and discussions of when and why people are influenced by false and misleading information. First, across chapters, we consistently find evidence indicating that people may not engage in critical evaluation of the accuracy of information during routine comprehension experiences, with detrimental downstream consequences for what people believe or report to be true. Without instructions and/or prompts to evaluate, people exhibit an increased risk of reproducing false ideas — even in cases when the inaccuracies blatantly contradict people's correct prior knowledge. This suggests that people do not always leverage their accurate prior knowledge in-the-moment to detect and discount false information, particularly when tasks involve non-evaluative goals or participants are holding non-evaluative mindsets, such as rating the interest of presented ideas or indicating whether they would like a post on social media.

Fortunately, we also demonstrate that people are amenable to instructions, prompting, and nudges that encourage them to adopt evaluative mindsets. These mindsets can protect against the influence of false information. Participants who considered the accuracy of potentially false information were not only less likely to reproduce incorrect ideas from previously encountered inaccuracies, but also more likely to rely on their correct prior knowledge if and when it was available. This suggests that evaluation can encourage activation, availability, and the use of

existing knowledge to counter the effects of exposures to false information. These benefits were observed for not only the specific information that people were tasked with judging for accuracy, but also for other information presented in the same experience, and even when that information was explicitly associated with non-evaluative tasks. This foregrounds the crucial role of evaluation for supporting comprehension, and highlights when and how evaluation can be usefully leveraged to reduce the influence of false information.

We also successfully tested the generalizability of these findings by applying them to a social media context, examining people's use of inaccuracies encountered on Twitter. Consistent with previous work that has largely leveraged decontextualized statements of true and false ideas, we found that accuracy mindsets reduced participants' reproductions of false ideas from tweets, and increased their production of correct answers based on accurate knowledge. These benefits emerged for participants who were explicitly told to evaluate the accuracy of the tweets, and also for participants who spontaneously reported accuracy considerations for making decisions about whether the contents of the tweet were interesting, or whether they would like the post if they encountered it on Twitter. (To be clear, people prompted to engage in interest and social media judgments reported similarly low rates of accuracy considerations in making their initial evaluations, and neither interest nor propensity to like a tweet predicted later reliance on the information.) In sum, people may not routinely pay attention to accuracy when engaging with information online, but are amenable to prompts that encourage an accuracy focus. That accuracy focus can protect against the influence of false information presented on social media.

Future investigations should expand this work to other contexts and environments, including social media posts and platforms that afford information consumption in a variety of modalities (e.g., short form videos on TikTok, or infographics prevalent on Instagram), and in

other everyday discourse experiences where inaccuracies may be communicated (e.g., via conversation). Preliminary results from recent work conducted in our lab suggests that evaluative mindsets can help reduce the extent to which people are influenced by incorrect answers offered by a partner during collaborative learning experiences, speaking to the generalizability of these interventions across contexts (Salovich, DeBode, Rapp, in prep). It would be both theoretically and practically informative to also test the efficacy of accuracy evaluations for reducing the influence of a variety of different kinds of falsehoods, perhaps ranging in familiarity, consequentiality, and/or ideological importance (e.g., Rapp et al., 2020). The experiments in this dissertation leveraged declarative general knowledge statements that have agreed upon and evidence-based true or false answers. While this is useful for achieving experimental control over people's prior knowledge about presented ideas, it does not represent the range of information that people engage with on a daily basis. It is crucial to investigate the consequences of evaluation using a wider array of topics that reflect the socio-cultural, emotional, and complexity of prior beliefs and understandings (e.g., anti-vax rhetoric, political misinformation), especially since the current dissertation underscores the importance of prior knowledge in the processes and benefits associated with evaluative considerations. Future projects should also explore ways to encourage evaluative strategies with different ideas using external resources or lateral reading, which may decrease reliance on inaccuracies that people may have difficulty evaluating based on prior knowledge alone (e.g., Donovan & Rapp, 2020; Wineburg & McGrew, 2019). Existing and future interventions would also benefit from verification using techniques that combine behavioral measures of belief with social media log data (e.g., clicks and shares; Pennycook & Rand, 2022), to support the goal of establishing potential long-term positive

effects of evaluative mindsets in real-world settings (Carey et al., 2022; Roozenbeek et al., 2021).

Research dedicated to understanding the consequences of exposure to inaccurate information has crucial implications for people's experiences with fake news, misinformation, developing stories, and misleading or sensationalist content that is becoming ubiquitous in online ecosystems. This dissertation was intended to contribute to that literature and current accounts that examine in-the-moment representations and downstream effects of experiences with inaccuracies. It also offers suggestions that can be applied in real-world applications to combat the spread of and belief in false information.

References

- Albrecht, J. E., & O'Brien, E. J. (1993). Updating a mental model: Maintaining both local and global coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1061–1070. <https://doi.org/10.1037/0278-7393.19.5.1061>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. <https://doi.org/10.1257/jep.31.2.211>
- Altay, S., Hacquin, A.-S., & Mercier, H. (2020). Why do so few people share fake news? It hurts their reputation. *New Media & Society*, 1461444820969893. <https://doi.org/10.1177/1461444820969893>
- Anderson, J. R. (1981). Effects of prior knowledge on memory for new information. *Memory & Cognition*, *9*(3), 237–246. <https://doi.org/10.3758/BF03196958>
- Andrews-Todd, J., Salovich, N. A., & Rapp, D. N. (2021). Differential effects of pressure on social contagion of memory. *Journal of Experimental Psychology: Applied*, *27*(2), 258–275. <https://doi.org/10.1037/xap0000346>
- Andrews, J. J., & Rapp, D. N. (2014). Partner characteristics and social contagion: Does group composition matter? *Applied Cognitive Psychology*, *28*(4), 505–517. <https://doi.org/10.1002/acp.3024>
- Andrews, J. J., & Rapp, D. N. (2015). Benefits, costs, and challenges of collaboration for learning and memory. *Translational Issues in Psychological Science*, *1*(2), 182–191. <https://doi.org/10.1037/tps0000025>
- Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, *27*(6), 576–605. [https://doi.org/10.1016/0022-1031\(91\)90026-3](https://doi.org/10.1016/0022-1031(91)90026-3)
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. <https://doi.org/10.1037/xge0000729>

- Baker, L., & Wagner, J. L. (1987). Evaluating information for truthfulness: The effects of logical subordination. *Memory & Cognition*, *15*(3), 247–255. <https://doi.org/10.3758/BF03197723>
- Bates, D., Mächler, M., Bolker, B., Walker, S., Christensen, R. H., & Singmann, H. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.
- Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*(4), 446–458. <https://doi.org/10.1037/0096-3445.121.4.446>
- Belli, R. F. (1989). Influences of misleading postevent information: Misinformation interference and acceptance. *Journal of Experimental Psychology: General*, *118*(1), 72–85. <https://doi.org/10.1037/0096-3445.118.1.72>
- Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In *Practical aspects of memory: Current research and issues, Vol. 1: Memory in everyday life* (pp. 396–401). John Wiley & Sons.
- Bond, C. F., & Titus, L. J. (1983). Social facilitation: A meta-analysis of 241 studies. *Psychological Bulletin*, *94*(2), 265–292. <https://doi.org/10.1037/0033-2909.94.2.265>
- Braasch, J. L. G., & Bråten, I. (2017). The discrepancy-induced source comprehension (D-ISC) model: Basic assumptions and preliminary evidence. *Educational Psychologist*, *52*(3), 167–181. <https://doi.org/10.1080/00461520.2017.1323219>
- Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515. <https://doi.org/10.1146/annurev-psych-010419-050807>
- Brashier, N. M., & Schacter, D. L. (2020). Aging in an era of fake news. *Current Directions in Psychological Science*, *29*(3), 316–323. <https://doi.org/10.1177/0963721420915872>
- Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*, 104054. <https://doi.org/10.1016/j.cognition.2019.104054>

- Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), e2020043118.
<https://doi.org/10.1073/pnas.2020043118>
- Britt, M. A., Rouet, J.-F., Blaum, D., & Millis, K. (2019). A reasoned approach to dealing with fake news. *Policy Insights from the Behavioral and Brain Sciences*, *6*(1), 94–101.
<https://doi.org/10.1177/2372732218814855>
- Butler, A. C., Dennis, N. A., & Marsh, E. J. (2012). Inferring facts from fiction: Reading correct and incorrect information affects memory for related information. *Memory (Hove, England)*, *20*(5), 487–498.
<https://doi.org/10.1080/09658211.2012.682067>
- Calvillo, D. P., & Smelter, T. J. (2020). An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications*, *5*(1), 55.
<https://doi.org/10.1186/s41235-020-00257-y>
- Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B., & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature Human Behaviour*, *6*(2), 236–243. <https://doi.org/10.1038/s41562-021-01278-3>
- Carrieri, V., Madio, L., & Principe, F. (2019). Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy. *Health Economics*, *28*(11), 1377–1382. <https://doi.org/10.1002/hec.3937>
- Chen, O., Paas, F., & Sweller, J. (2021). Spacing and interleaving effects require distinct theoretical bases: A systematic review testing the cognitive load and discriminative-contrast hypotheses. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-021-09613-w>
- Clark, H. H. (1977). Bridging. In P. N. Johnson-Laird, & P. C. Wason (Eds.), *Thinking: Readings in Cognitive Science* (pp. 411-420). Cambridge: Cambridge University Press.
- Cook, A. E., & O'Brien, E. J. (2014). Knowledge activation, integration, and validation during narrative text comprehension. *Discourse Processes*, *51*(1–2), 26–49. <https://doi.org/10.1080/0163853X.2013.855107>

- Cook, A. E., Walsh, E. K., Bills, M. A. A., Kircher, J. C., & O'Brien, E. J. (2018). Validation of semantic illusions independent of anomaly detection: Evidence from eye movements. *Quarterly Journal of Experimental Psychology*, *71*(1), 113–121. <https://doi.org/10.1080/17470218.2016.1264432>
- Corneille, O., Mierop, A., & Unkelbach, C. (2020). Repetition increases both the perceived truth and fakeness of information: An ecological account. *Cognition*, *205*, 104470. <https://doi.org/10.1016/j.cognition.2020.104470>
- Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*(2), 238–257. <https://doi.org/10.1177/1088868309352251>
- Dewey, C. (2016). Facebook fake-news writer: 'I think Donald Trump is in the White House because of me.' *Washington Post*. Retrieved November 28, 2021, from <https://www.washingtonpost.com/news/the-intersect/wp/2016/11/17/facebook-fake-news-writer-i-think-donald-trump-is-in-the-white-house-because-of-me/>
- Donovan, A. M., & Rapp, D. (2020). Look it up: Online search reduces the problematic effects of exposures to inaccuracies. *Memory & Cognition*. <https://doi.org/10.3758/s13421-020-01047-z>
- 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. <https://doi.org/10.3758/MC.38.8.1087>
- Effron, D. A. (2018). It could have been true: How counterfactual thoughts reduce condemnation of falsehoods and increase political polarization. *Personality and Social Psychology Bulletin*, *44*(5), 729–745. <https://doi.org/10.1177/0146167217746152>
- Epstein, Z., Sirlin, N., Arechar, A. A., Pennycook, G., & Rand, D. (2021). *Social media sharing reduces truth discernment*. <https://doi.org/10.31234/osf.io/q4bd2>
- Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning & Verbal Behavior*, *20*(5), 540–551. [https://doi.org/10.1016/S0022-5371\(81\)90165-1](https://doi.org/10.1016/S0022-5371(81)90165-1)

- Ernala, S. K., Burke, M., Leavitt, A., & Ellison, N. B. (2020). How well do people report time spent on Facebook? An evaluation of established survey questions with recommendations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.
<https://doi.org/10.1145/3313831.3376435>
- Eslick, A. N., Fazio, L. K., & Marsh, E. J. (2011). Ironic effects of drawing attention to story errors. *Memory*, 19(2), 184–191. <https://doi.org/10.1080/09658211.2010.543908>
- Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, 1(2). <https://doi.org/10.37016/mr-2020-009>
- Fazio, L. K., & Marsh, E. J. (2008). Slowing presentation speed increases illusions of knowledge. *Psychonomic Bulletin & Review*, 15(1), 180–185. <https://doi.org/10.3758/PBR.15.1.180>
- Fazio, L. K., Barber, S. J., Rajaram, S., Ornstein, P. A., & Marsh, E. J. (2013). Creating illusions of knowledge: Learning errors that contradict prior knowledge. *Journal of Experimental Psychology: General*, 142(1), 1–5. <https://doi.org/10.1037/a0028649>
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5), 993–1002.
<https://doi.org/10.1037/xge0000098>
- Fazio, L. K., Dolan, P. O., & Marsh, E. J. (2015). Learning misinformation from fictional sources: Understanding the contributions of transportation and item-specific processing. *Memory*, 23(2), 167–177.
<https://doi.org/10.1080/09658211.2013.877146>
- Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, 26(5), 1705–1710.
<https://doi.org/10.3758/s13423-019-01651-4>
- Fife, D. (2014). *fifer*: A collection of miscellaneous functions. R package version 1.0.
- French, L., Garry, M., & Mori, K. (2011). Relative–not absolute–judgments of credibility affect susceptibility to misinformation conveyed during discussion. *Acta Psychologica*, 136(1), 119–128.
<https://doi.org/10.1016/j.actpsy.2010.10.009>

- Gerrig, R. J., & Prentice, D. A. (1991). The representation of fictional information. *Psychological Science*, 2(5), 336–340. <https://doi.org/10.1111/j.1467-9280.1991.tb00162.x>
- Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, 46(2), 107–119. <https://doi.org/10.1037/0003-066X.46.2.107>
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59(4), 601–613. <https://doi.org/10.1037/0022-3514.59.4.601>
- Grice, H. P. (1975). Logic and conversation. *Speech Acts*, 41–58. https://doi.org/10.1163/9789004368811_003
- Groggel, A., Nilizadeh, S., Ahn, Y.-Y., Kapadia, A., & Rojas, F. (2019). Race and the beauty premium: Mechanical Turk workers' evaluations of Twitter accounts. *Information, Communication & Society*, 22(5), 709–716. <https://doi.org/10.1080/1369118X.2018.1543443>
- Guess, A., Nagler, J., & Tucker, J. (n.d.). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, 5(1), eaau4586. <https://doi.org/10.1126/sciadv.aau4586>
- Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669), 438–441. <https://doi.org/10.1126/science.1095455>
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning & Verbal Behavior*, 16(1), 107–112. [https://doi.org/10.1016/S0022-5371\(77\)80012-1](https://doi.org/10.1016/S0022-5371(77)80012-1)
- Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, 6(1), 38. <https://doi.org/10.1186/s41235-021-00301-5>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, 19(2), 212–225. <https://doi.org/10.1086/209297>

- Hawkins, S. A., Hoch, S. J., & Meyers-Levy, J. (2001). Low-involvement learning: Repetition and coherence in familiarity and belief. *Journal of Consumer Psychology, 11*(1), 1–11.
https://doi.org/10.1207/S15327663JCP1101_1
- Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition, 32*(2), 336–350. <https://doi.org/10.3758/BF03196863>
- Hinze, S. R., Slaten, D. G., Horton, W. S., Jenkins, R., & Rapp, D. N. (2014). Pilgrims sailing the Titanic: Plausibility effects on memory for misinformation. *Memory & Cognition, 42*(2), 305–324.
<https://doi.org/10.3758/s13421-013-0359-9>
- Indicators of news media trust.* (2018). Knight Foundation. Retrieved November 28, 2021, from
<https://knightfoundation.org/reports/indicators-of-news-media-trust/>
- Isberner, M.-B., & Richter, T. (2014a). Comprehension and validation: Separable stages of information processing? A case for epistemic monitoring in language comprehension. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 245–276). Boston Review.
- Isberner, M.-B., & Richter, T. (2014b). Does validation during language comprehension depend on an evaluative mindset? *Discourse Processes, 51*(1–2), 7–25. <https://doi.org/10.1080/0163853X.2013.855867>
- Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive reminders in recency judgments and cued recall. *Memory & Cognition, 41*(5), 625–637.
<https://doi.org/10.3758/s13421-013-0298-5>
- Jalbert, M., Newman, E., & Schwarz, N. (2020). Only half of what I'll tell you is true: Expecting to encounter falsehoods reduces illusory truth. *Journal of Applied Research in Memory and Cognition, 9*(4), 602–613.
<https://doi.org/10.1016/j.jarmac.2020.08.010>
- Kazdin, A. E. (1974). Reactive self-monitoring: The effects of response desirability, goal setting, and feedback. *Journal of Consulting and Clinical Psychology, 42*(5), 704–716. <https://doi.org/10.1037/h0037050>

- Kelley, C. M. ., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32(1), 1–24. <https://doi.org/10.1006/jmla.1993.1001>
- Kendeou, P., & O'Brien, E. J. (2014). The knowledge revision components (KReC) framework: Processes and mechanisms. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 353–377). Boston Review.
- Kitchner, K. S. (1983). Cognition, metacognition, and epistemic cognition. *Human Development*, 26(4), 222–232. <https://doi.org/10.1159/000272885>
- Van-Dijk, D., & Kluger, A. N. (2000). Positive (negative) feedback: Encouragement or discouragement. *The Annual Convention of the Society for Industrial and Organizational Psychology, New Orleans, Louisiana*.
- Van-Dijk, D., & Kluger, A. N. (2001). Goal orientation versus self-regulation: Different labels or different constructs. *16th annual convention of the Society for Industrial and Organizational Psychology, San Diego, CA*.
- Kumle, L., Vö, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, 53(6), 2528–2543. <https://doi.org/10.3758/s13428-021-01546-0>
- Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427), 203–205. <https://doi.org/10.1126/science.7350657>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82, 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>

- Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development*, 58(6), 629–648. <https://doi.org/10.1007/s11423-010-9153-6>
- Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Estimated marginal means, aka least-squares means. R package version 1.3. 2.
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the “Post-Truth” era. *Journal of Applied Research in Memory and Cognition*, 6(4), 353–369. <https://doi.org/10.1016/j.jarmac.2017.07.008>
- Lewis, C. H., & Anderson, J. R. (1976). Interference with real world knowledge. *Cognitive Psychology*, 8(3), 311–335. [https://doi.org/10.1016/0010-0285\(76\)90010-4](https://doi.org/10.1016/0010-0285(76)90010-4)
- Loftus, E. F. (1979). Reactions to blatantly contradictory information. *Memory & Cognition*, 7(5), 368–374. <https://doi.org/10.3758/BF03196941>
- Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*, 5(3), 337–348. <https://doi.org/10.1038/s41562-021-01056-1>
- Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 49(4), 1494–1502. <https://doi.org/10.3758/s13428-016-0809-y>
- Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences*, 118(23). <https://doi.org/10.1073/pnas.2019527118>
- Marsh, E. J., & Fazio, L. K. (2006). Learning errors from fiction: Difficulties in reducing reliance on fictional stories. *Memory & Cognition*, 34(5), 1140–1149. <https://doi.org/10.3758/BF03193260>
- Marsh, E. J., Cantor, A. D., & M. Brashier, N. (2016). Believing that humans swallow spiders in their sleep: False beliefs as side effects of the processes that support accurate knowledge. In B. H. Ross (Ed.),

Psychology of Learning and Motivation (Vol. 64, pp. 93–132). Academic Press.

<https://doi.org/10.1016/bs.plm.2015.09.003>

Marsh, E. J., Meade, M. L., & Roediger III, H. L. (2003). Learning facts from fiction. *Journal of Memory and Language*, 49(4), 519–536. [https://doi.org/10.1016/S0749-596X\(03\)00092-5](https://doi.org/10.1016/S0749-596X(03)00092-5)

Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, 26(1), 283–327. <https://doi.org/10.1080/10463283.2015.1117249>

Mayo, R. (2019). The skeptical (ungullible) mindset. In *The Social Psychology of Gullibility*. Routledge.

Mayo, R., Alfasi, D., & Schwarz, N. (2014). Distrust and the positive test heuristic: Dispositional and situated social distrust improves performance on the Wason Rule Discovery Task. *Journal of Experimental Psychology: General*, 143(3), 985–990. <https://doi.org/10.1037/a0035127>

Mena, P., Barbe, D., & Chan-Olmsted, S. (2020). Misinformation on Instagram: The impact of trusted endorsements on message credibility. *Social Media + Society*, 6(2), 2056305120935102.

<https://doi.org/10.1177/2056305120935102>

Negley, J. H., Kelley, C. M., & Jacoby, L. L. (2018). The importance of time to think back: The role of reminding in retroactive effects of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(9), 1352–1364. <https://doi.org/10.1037/xlm0000512>

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2006). The effect of distributed monitoring exercises and feedback on performance, monitoring accuracy, and self-efficacy. *Metacognition and Learning*, 1(2), 159.

<https://doi.org/10.1007/s10409-006-9595-6>

O'Brien, E. J., & Cook, A. E. (2016a). Separating the activation, integration, and validation components of reading. In *The psychology of learning and motivation* (pp. 249–276). Elsevier Academic Press.

O'Brien, E. J., & Cook, A. E. (2016b). Coherence threshold and the continuity of processing: The RI-Val model of comprehension. *Discourse Processes*, 53(5–6), 326–338.

<https://doi.org/10.1080/0163853X.2015.1123341>

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science*, 3(4), 229–236. <https://doi.org/10.1111/j.1467-9280.1992.tb00034.x>

- Paris, S. G., & Winograd, P. (1990). Promoting metacognition and motivation of exceptional children. *Remedial and Special Education, 11*(6), 7–15. <https://doi.org/10.1177/074193259001100604>
- Parkinson, H. J. (2016, November 14). Click and elect: How fake news helped Donald Trump win a real election. *The Guardian*. <https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies>
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(1), 3–8. <https://doi.org/10.1037/0278-7393.31.1.3>
- Pasquetto, I. V., Swire-Thompson, B., Amazeen, M. A., Benevenuto, F., Brashier, N. M., Bond, R. M., Bozarth, L. C., Budak, C., Ecker, U. K. H., Fazio, L. K., Ferrara, E., Flanagan, A. J., Flammini, A., Freelon, D., Grinberg, N., Hertwig, R., Jamieson, K. H., Joseph, K., Jones, J. J., ... Yang, K.-C. (2020). Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*. <https://dash.harvard.edu/handle/1/37366685>
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25*(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pennycook, G., & Rand, D. G. (2022). Nudging social media toward accuracy. *The ANNALS of the American Academy of Political and Social Science, 700*(1), 152–164. <https://doi.org/10.1177/00027162221092342>
- Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology, 7*(1). <https://doi.org/10.1525/collabra.25293>
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595. <https://doi.org/10.1038/s41586-021-03344-2>

- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770–780. <https://doi.org/10.1177/0956797620939054>
- Podsakoff, P. M., & Farh, J.-L. (1989). Effects of feedback sign and credibility on goal setting and task performance. *Organizational Behavior and Human Decision Processes*, 44(1), 45–67. [https://doi.org/10.1016/0749-5978\(89\)90034-4](https://doi.org/10.1016/0749-5978(89)90034-4)
- Prentice, D. A., & Gerrig, R. J. (1999). Exploring the boundary between fiction and reality. In *Dual-process theories in social psychology* (pp. 529–546). The Guilford Press.
- Prentice, D. A., Gerrig, R. J., & Bailis, D. S. (1997). What readers bring to the processing of fictional texts. *Psychonomic Bulletin & Review*, 4(3), 416–420. <https://doi.org/10.3758/BF03210803>
- Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Memory & Cognition*, 36(3), 688–701. <https://doi.org/10.3758/MC.36.3.688>
- Rapp, D. N. (2016). The consequences of reading inaccurate information. *Current Directions in Psychological Science*, 25(4), 281–285. <https://doi.org/10.1177/0963721416649347>
- Rapp, D. N., & Braasch, J. L. G. (2014). *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. MIT Press.
- Rapp, D. N., & Mensink, M. C. (2011). Focusing effects from online and offline reading tasks. In *Text relevance and learning from text* (pp. 141–164). IAP Information Age Publishing.
- Rapp, D. N., & Salovich, N. A. (2018). Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences*, 5(2), 232–239. <https://doi.org/10.1177/2372732218785193>
- Rapp, D. N., Donovan, A. M., & Salovich, N. A. (2020). Assessing and modifying knowledge: Facts vs. constellations. In *Handbook of Learning from Multiple Representations and Perspectives*. Routledge.
- Rapp, D. N., Hinze, S. R., Kohlhepp, K., & Ryskin, R. A. (2014). Reducing reliance on inaccurate information. *Memory & Cognition*, 42(1), 11–26. <https://doi.org/10.3758/s13421-013-0339-0>

- Rapp, D. N., Jacovina, M. E., & Andrews, J. J. (2014). Mechanisms of problematic knowledge acquisition. In *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 181–202). Boston Review. <https://doi.org/10.7551/mitpress/9737.003.0013>
- Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, 1(4), 563–581. <https://doi.org/10.1007/s13164-010-0039-7>
- Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, 41(3), 221–250. https://doi.org/10.1207/s15326950dp4103_1
- Richter, T. (2015). Validation and comprehension of text information: Two sides of the same coin. *Discourse Processes*, 52(5–6), 337–355. <https://doi.org/10.1080/0163853X.2015.1025665>
- Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology*, 96(3), 538–558. <https://doi.org/10.1037/a0014038>
- Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Roll, I., Alevin, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267–280. <https://doi.org/10.1016/j.learninstruc.2010.07.004>
- Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A reregistered direct replication of Pennycook et al. (2020). *Psychological Science*, 32(7), 1169–1178. <https://doi.org/10.1177/09567976211024535>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>
- Salovich, N. A., DeBode, E. G., & Rapp, D. N. (2022). The social contagion of knowledge: Do people reproduce others' incorrect answers? Manuscript in prep.

- Salovich, N. A., Donovan, A. M., Hinze, S. R., & Rapp, D. N. (2021). Can confidence help account for and redress the effects of reading inaccurate information? *Memory & Cognition*, *49*(2), 293–310.
<https://doi.org/10.3758/s13421-020-01096-4>
- Salovich, N. A., Imundo, M. N., & Rapp, D. N. (2022). Story stimuli for instantiating true and false beliefs about the world. *Behavioral Research Methods*, in press.
- Salovich, N. A., Kirsch, A. M., & Rapp, D. N. (2022). Evaluative mindsets can protect against the influence of false information. *Cognition*, *225*, 105121. <https://doi.org/10.1016/j.cognition.2022.105121>
- Salovich, N. A., & Rapp, D. N. (2021). Misinformed and unaware? Metacognition and the influence of inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(4), 608–624. <https://doi.org/10.1037/xlm0000977>
- Sanford, A. J., Leuthold, H., Bohan, J., & Sanford, A. J. S. (2011). Anomalies at the borderline of awareness: An ERP study. *Journal of Cognitive Neuroscience*, *23*(3), 514–523.
<https://doi.org/10.1162/jocn.2009.21370>
- Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, *7*(4), 351–371.
<https://doi.org/10.1007/BF02212307>
- Schul, Y., Mayo, R., & Burnstein, E. (2004). Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *Journal of Personality and Social Psychology*, *86*(5), 668–679.
<https://doi.org/10.1037/0022-3514.86.5.668>
- Shoemaker, P. J., & Vos, T. (2009). *Gatekeeping Theory*. Routledge. <https://doi.org/10.4324/9780203931653>
- Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language*, *54*(4), 574–591. <https://doi.org/10.1016/j.jml.2005.11.003>
- Singer, M. (2013). Validation in reading comprehension. *Current Directions in Psychological Science*, *22*(5), 361–366.
- Singer, M. (2019). Challenges in processes of validation and comprehension. *Discourse Processes*, *56*(5–6), 465–483. <https://doi.org/10.1080/0163853X.2019.1598167>

- Sparks, J. R., & Rapp, D. N. (2011). Readers' reliance on source credibility in the service of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(1), 230–247. <https://doi.org/10.1037/a0021331>
- Storm, B., C. (2011). Retrieval-induced forgetting and the resolution of competition. In *Successful Remembering and Successful Forgetting*. Psychology Press.
- Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods*, 45(4), 1115–1143. <https://doi.org/10.3758/s13428-012-0307-9>
- Thebault, R. (2019). Facebook says it will take action against anti-vaccine content. Here's how it plans to do it. *Washington Post*. Retrieved November 28, 2021, from <https://www.washingtonpost.com/business/2019/03/07/facebook-says-it-will-take-action-against-anti-vax-content-heres-how-they-plan-do-it/>
- Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). *Social media, political polarization, and political disinformation: A review of the scientific literature* (SSRN Scholarly Paper ID 3144139). Social Science Research Network. <https://doi.org/10.2139/ssrn.3144139>
- van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*, 39(2–3), 299–316. https://doi.org/10.1207/s15326950dp3902&3_11
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory & Cognition*, 43(1), 27–38. <https://doi.org/10.3758/s13421-014-0455-5>
- Wahlheim, C. N., & Jacoby, L. L. (2013). Remembering change: The critical role of recursive reminders in proactive effects of memory. *Memory & Cognition*, 41(1), 1–15. <https://doi.org/10.3758/s13421-012-0246-9>

- Walker, M., & Matsa, K. E. (2021, September 20). News consumption across social media in 2021. *Pew Research Center's Journalism Project*. <https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/>
- Waruwu, B. K., Tandoc, E. C., Duffy, A., Kim, N., & Ling, R. (2021). Telling lies together? Sharing news as a form of social authentication. *New Media & Society*, 23(9), 2516–2533. <https://doi.org/10.1177/1461444820931017>
- Weil, R., & Mudrik, L. (2020). Detecting falsehood relies on mismatch detection between sentence components. *Cognition*, 195, 104121. <https://doi.org/10.1016/j.cognition.2019.104121>
- Weil, R., Schul, Y., & Mayo, R. (2020). Correction of evident falsehood requires explicit negation. *Journal of Experimental Psychology: General*, 149(2), 290–310. <https://doi.org/10.1037/xge0000635>
- Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*, 121(11), 1–40. <https://doi.org/10.1177/016146811912101102>
- Wiswede, D., Koranyi, N., Müller, F., Langner, O., & Rothermund, K. (2013). Validating the truth of propositions: Behavioral and ERP indicators of truth evaluation processes. *Social Cognitive and Affective Neuroscience*, 8(6), 647–653. <https://doi.org/10.1093/scan/nss042>
- Zawadzka, K., Krogulska, A., Button, R., Higham, P. A., & Hanczakowski, M. (2016). Memory, metamemory, and social cues: Between conformity and resistance. *Journal of Experimental Psychology: General*, 145(2), 181–199. <https://doi.org/10.1037/xge0000118>

Appendix A

Positive Interim Feedback

Qualtrics has now finished scoring your responses.

As you may have noticed, some of the statements that you had previously read were false. **The scoring system indicates that, based on others' prior performance on this task, you reproduced fewer inaccurate answers from those statements on the questionnaire than the average Northwestern student.** In other words, your responses were minimally influenced by the false information that you viewed.

In the second half of the experiment, you will view a new set of statements and answer a new set of questions. Please continue to begin.

Negative Interim Feedback

Qualtrics has now finished scoring your responses.

As you may have noticed, some of the statements that you had previously read were false. **The scoring system indicates that, based on others' prior performance on this task, you reproduced more inaccurate answers from those statements on the questionnaire than the average Northwestern student.** In other words, your responses were highly influenced by the false information that you viewed.

In the second half of the experiment, you will view a new set of statements and answer a new set of questions. Please continue to begin.

No Interim Feedback/Control


Qualtrics has now finished scoring your responses.

In the second half of the experiment, you will view a new set of statements and answer a new set of questions. Please continue to begin.

Appendix B




Easy Item Example (Item #8)

True version

 **Sammy Ellis** @s_r_ellis ⋮

My daughter won first place in her track meet today!
She is as fast as a cheetah — the fastest land animal on Earth.

4:36 PM · Mar 6, 2021

False version

 **Sammy Ellis** @s_r_ellis ⋮

My daughter won first place in her track meet today!
She is as fast as a horse — the fastest land animal on Earth.

4:36 PM · Mar 6, 2021

Filler version

 **Sammy Ellis** @s_r_ellis ⋮

My daughter won first place in her track meet today!
She is as fast as the fastest land animal on Earth.

4:36 PM · Mar 6, 2021

Hard Item Example (Item #59)*True version*

Harper Morgan
@harper_m101



On our cruise, the captain mentioned we were passing Napoleon's birthplace in Corsica. That makes this count as an educational trip, right?

12:04 PM · Jul 26, 2021

*False version*

Harper Morgan
@harper_m101



On our cruise, the captain mentioned we were passing Napoleon's birthplace in Tahiti. That makes this count as an educational trip, right?

12:04 PM · Jul 26, 2021

*Filler version*

Harper Morgan
@harper_m101



On our cruise, the captain mentioned we were passing Napoleon's birthplace. That makes this count as an educational trip, right?

12:04 PM · Jul 26, 2021

