Journal of Applied Research in Memory and Cognition 9 (2020) 602–613

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Empirical article

Only Half of What I'll Tell You is True: Expecting to Encounter Falsehoods Reduces Illusory Truth

Madeline Jalbert* and Norbert Schwarz
University of Southern California, United States

Eryn Newman
Australian National University, Australia

Information is judged as more true when it has been seen or heard repeatedly than when it is new. This *illusory truth effect* has important consequences in the real world, where we are repeatedly exposed to information of unknown veracity. While false information in natural contexts rarely comes with a warning label, false information in truth effect experiments often does. Commonly used experimental procedures alert participants to potential falsehoods at exposure through instructional warnings. Three experiments show that the size of the truth effect is over twice as large when such warnings are avoided. The influence of pre-exposure warnings on the size of the truth effect persists even after a delay of three to six days. These findings demonstrate that common experimental procedures invite a systematic underestimation of illusory truth effects. They also highlight that simple warnings can curb the impact of repetition on judgments of truth.

Keywords: Truth effect, Fluency, Metacognition

General Audience Summary

People are more likely to believe a statement when they have seen or heard it before, a phenomenon called the *illusory truth effect*. This has important implications for daily life, where we are repeatedly exposed to both true and false information as we scroll through social media, read the news, or talk with others. We test whether the influence of repetition on belief depends on whether one is warned that information presented may be false. In three experiments, we first asked participants to read a series of trivia claims. Half of the claims were true and half were false. We explicitly told some of the participants that some claims were false, whereas other participants were not alerted to this. After a delay, participants saw another set of trivia claims, including ones they had already seen before and ones that were new. As in earlier studies, participants believed the repeated claims were more true than claims they read for the first time. Importantly, the influence of repetition on belief was over twice as large when participants had not been warned that some claims were false. The protective effect of these initial warnings was observed even when participants did not judge the truth of those claims until three to six days later. However, the warnings were only helpful when they preceded the first reading of the

Author Note

Eryn Newman, Australian National University, Australia.

Norbert Schwarz, Department of Psychology, University of Southern California, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061, USA.

Preparation of this article was supported by the Linnie and Michael Katz Endowed Research Fellowship Fund through a fellowship to the first author.

* Correspondence concerning this article should be addressed to Madeline Jalbert, Department of Psychology, University of Southern California, 3620 S. McClintock Ave, Los Angeles, CA 90089-1061, USA. Contact: mcjalber@usc.edu.

claims. Waiting to warn people until they later had to judge their truth had no detectable influence. These results show that warnings can curb the influence of repetition on belief in false information, provided the warning precedes initial exposure. They also show that many truth effect experiments may have underestimated the impact of repetition on belief due to the presence of warnings in their experimental designs that are usually absent under real world conditions.

For millennia, demagogues believed that repetition can turn lies into truth. Experimental research confirmed their intuitions. Four decades ago, Hasher, Goldstein, & Toppino (1977) reported that the mere repetition of a statement increased its acceptance as true. This so-called *illusory truth effect* proved robust and easily replicable, with a meta-analysis of seventy effect sizes reporting an average between-items effect size of $d = 0.50$, 95% CI [0.43, 0.57], for the difference in truth ratings between new and repeated items (Dechêne, Stahl, Hansen, & Wänke, 2010). Current concerns about the prevalence of fake news and their repetition on social media have renewed interest in the issue. Unfortunately, the standard procedures of truth effect experiments may not be a good approximation of the conditions of message repetition in natural contexts. Most commonly used experimental procedures create exposure conditions that draw attention to the fact that some of the information one will see is false. Such pre-exposure warnings may systematically decrease the impact of repetition compared to natural conditions, where false information does not come with a warning label. We test this possibility by investigating the size of illusory truth effects under conditions that do or do not include common pre-exposure warnings.

Judgment and persuasion research suggests that people draw on a limited set of criteria to determine whether a claim is likely to be true (Schwarz, 2015, 2018): Is the claim compatible with other things they believe? Is it internally consistent? Does it come from a credible source? Do others believe it? Is there supporting evidence? Each of these truth criteria can be assessed by careful attention to the claim and reliance on applicable knowledge (for reviews, see Brashier & Marsh, 2020; Schwarz, 2015; Unkelbach, Koch, Silva, & Garcia-Marques, 2019). They can also be assessed by drawing on one's metacognitive experiences as a proxy. In each case, the attribute that provides an affirmative answer to the criterion is correlated with fluent processing. For example, when information is compatible with other things one believes (Unkelbach & Rom, 2017; Winkielman, Huber, Kavanagh, & Schwarz, 2012), internally coherent (Johnson-Laird, 2012), or familiar from previous exposures (Jacoby, 1983), it is processed more fluently than when it is not. Moreover, familiar sources are more credible (Petty & Cacioppo, 1986) and supporting evidence is assumed to be more plentiful when some comes to mind easily (Schwarz, 1998; Tversky & Kahneman, 1973). This makes the metacognitive experience of processing fluency a heuristically informative input into judgments of truth, independent of whether people assess the claim's compatibility with their own knowledge, its coherence, social consensus, source credibility, or the likely amount of supporting evidence. Indeed, fluent processing can override one's own knowledge; when information feels true and no alternative accounts easily come to mind, people may accept it without performing the more effortful analysis that would lead them to find it false (Brashier, Eliseev, & Marsh, 2020; for a review, see Brashier & Marsh, 2020).

That processing fluency is correlated with substantive attributes relevant to judging truth implies that it provides valid information. Unfortunately, people are more sensitive to their processing experience itself than to the source of this experience (Schwarz, 2010) and sometimes misread fluent processing due to incidental influences as bearing on the truth of a statement. Indeed, numerous incidental variables that affect processing fluency have been found to influence judgments of truth, from print font and color contrast (e.g., Garcia-Marques, Silva, & Mello, 2016; Parks & Toth, 2006; Reber & Schwarz, 1999; Silva, Garcia-Marques, & Mello, 2016) to accent (Lev-Ari & Keysar, 2010) and audio quality (Newman & Schwarz, 2018), to the ease of pronouncing the information's source (Newman et al., 2014). Such fluency effects are incidental and unrelated to the semantic content of the claim.

Repetition of a claim can influence judgments of truth through several pathways, namely increased perceptual and conceptual fluency, increased accessibility of applicable knowledge, and recollection of source information. Accordingly, variants of recollective and fluency-based process accounts have been offered since Hasher et al. (1977)'s experiments. Empirically, knowledge gleaned from recalling the context of prior exposure as well as processing experience contribute to repetition effects (Unkelbach & Stahl, 2009). Additionally, the impact of repetition often exceeds that of perceptual fluency manipulations, such as print font (e.g., Parks & Toth, 2006), and strategies that are effective in attenuating the influence of perceptual fluency, such as stressing the need to be accurate when assessing truth, are less effective in correcting repetition-based truth effects (e.g., Garcia-Marques et al., 2016; Silva et al., 2016).

Pre-Exposure and Pre-Test Warnings

A typical investigation of repetition-based truth effects begins with an exposure phase, where participants view a series of ambiguous claims. Usually, half of these claims are true and half are false. Following a delay (ranging from a few minutes to several days), there is a test phase, where participants view claims

they saw during the exposure phase along with new claims and rate the truth of each claim.

Studies vary in the extent to which they draw participants' attention to the truth of the claims at initial exposure. In the majority of studies, researchers alert participants to the presence of potential falsehoods at exposure in at least one of two ways: by stating in the instructions that claims vary in truth value (e.g., Begg, Anas, & Farinacci, 1992; Hasher et al., 1977) or by asking participants to make a truth judgment for each claim as it is presented (e.g., Brown & Nix, 1996; Hasher et al., 1977). Occasionally, claims are explicitly labeled as true or false as they are presented (e.g., Skurnik, Yoon, Park, & Schwarz, 2005). In a small number of studies, participants are not alerted to the presence of potential falsehoods at exposure. Instead, participants may be asked about attributes other than truth (e.g., Brashier et al., 2020; Hawkins & Hoch, 1992) or simply view the claims for a later memory test (Mitchell, Sullivan, Schacter, & Budson, 2006).

At the time of testing, all truth effect studies draw participants' attention to the truth of the claims simply by asking participants to judge their truth. At this stage, some studies additionally provide participants with explicit instructional information alerting them that "some" or "half" of the claims are false (e.g., Begg et al., 1992; Brashier et al., 2020).

How might drawing attention to the presence of falsehoods at the time of encoding, either by asking participants to make truth judgments or through instructional details, influence the size of the truth effect? Previous research showed that making truth judgments during initial exposure reduces the impact of repetition (e.g., Hawkins & Hoch, 1992). More recent research focusing on false claims indicates that this protective influence may only occur for false claims for which participants have knowledge that allows them to arrive at the correct answer (Brashier et al., 2020). In a meta-analysis, Dechêne et al. (2010) obtained an effect size of d = 0.45, 95% CI [0.37, 0.54], for studies that included truth judgments at exposure as compared to d = 0.62, 95% CI [0.49, 0.75], for studies that asked for other judgments or no judgments at all. Dechêne et al. (2010) attributed this observation to differences in level of processing, but it is worth noting that asking for an assessment of truth amounts to acknowledging that truth cannot be taken for granted. From this perspective, the meta-analysis suggests that drawing attention to truth at encoding by requesting explicit judgments for each claim can reduce the size of the truth effect.

However, it is unclear whether merely informing participants that some statements will be false through standard experimental instructions can do the same thing. This question is difficult to answer meta-analytically because most of the available studies (51 out of 70 effect sizes included in Dechêne et al., 2010) include truth judgments at exposure, making the impact of instructional warnings difficult to isolate. Additionally, the methods sections of some reports lack the procedural details that would be necessary to determine if instructional warnings were given at the exposure stage.

Research on the encoding and correction of misinformation suggests that instructional warnings can influence later belief in false information (for a review, see Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). In general, people are better able to protect themselves from misinformation if a warning makes them skeptical of the accuracy of that information during encoding. Increased skepticism promotes more critical processing and reduces the acceptance of new information as true (e.g., Fein, McCloskey, & Tomlinson, 1997; Greene, Flynn, & Loftus, 1982; Lewandowsky, Stritzke, Oberauer, & Morales, 2005; Schul, 1993). Pre-exposure warnings that some statements are false may similarly elicit more critical processing at encoding, reducing the impact of repetition in truth effect experiments. These findings suggest that without instructional warnings at exposure, the truth effect may be larger than previously thought. Hence, the most frequently used experimental procedures may underestimate the impact of repetition under natural conditions.

It is less clear how instructional warnings at the time of test may influence the size of the truth effect. Being asked to make a truth judgment already entails that the claims are likely to vary in veracity. Hence, additional information that merely warns recipients that some of the claims are false may have little impact. Consistent with this assumption, Nadarevic and Aßfalg (2017) reported that only a detailed explanation of the truth effect and specific instructions for how to prevent it prior to test reduced the impact of repetition, whereas more general warnings did not. We similarly expect to find little impact of warnings presented at the time of testing, in contrast to warnings presented at the time of exposure.

Present Research

We investigated how the presence, timing, and content of instructional warnings influence the size of the truth effect. In Experiment 1, we manipulated the presence of pre-exposure warnings: participants either did or did not receive a warning that half of the claims are false before they were exposed to the claims. In Experiment 2, we tested the impact of pre-exposure and pre-test warnings: participants received a warning prior to exposure and test, a warning prior to test only, or no warning at all. This allowed us to determine whether instructional warnings at the time of testing have a protective value either by themselves or in combination with pre-test warnings. In Experiment 3, we varied the warnings by telling participants either that "some" or "half" of the statements would be false.

Experiment 1

The purpose of Experiment 1 was to investigate how the presence of standard instructional warnings prior to the initial exposure influences the size of the truth effect. Based on the observation that pre-exposure warnings reduce the impact of misleading information in other paradigms (e.g., Fein et al., 1997; Greene et al., 1982; Lewandowsky et al., 2005; Schul, 1993) we predicted that the impact of repetition is smaller when participants are aware, prior to exposure, that some information they are about to see may be false. We included a delay of three to six days between initial exposure and test to approximate the often long delays between exposure and re-exposure under natural conditions. We predicted that, compared to a no warning

condition, pre-exposure warnings would attenuate the impact of repetition on rated truth even after a multi-day delay. All stimuli, instructions, recruitment and attrition information, data, syntax, and additional analysis are included in the supplemental materials available at https://doi.org/10.3886/E115141V2.

Method

Design. We used a 2 (warning: before exposure only vs. no warning) × 2 (repetition: trivia claim repeated vs. new) mixed design, manipulating warning between subjects and repetition within subjects. The delay between exposure and test was 72 to 144 hours (three to six days).

Participants. Students from the University of Southern California psychology subject pool completed the survey for course credit. Data collection took place in two waves. Based on the between-items effects size of $d = 0.50$, 95% CI [0.43, 0.57], reported by Dechêne et al. (2010), a sample size of 54 would be required to detect the truth effect in a repeated measures design, with $\alpha = .05$, power $(1\text{-}\beta) = .95$, and two-tailed, according to G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). We chose to overpower our study and recruit up to 200 participants, 100 per between-subjects condition. However, we ended data collection when the subject pool closed and included all participants that had completed the experiment at that point. Unfortunately, by the time the subject pool closed, we had only recruited approximately one third of our desired number of participants. We therefore collected additional data in a second wave the following year, where we again opened the study up to 200 participants.

We include all participants in the analyses below, except those who did not complete part one ($n = 13$) within 24 hours of when they began the survey or part two ($n = 5$) within 48 hours after the email invitation for that part; both exclusions are necessary to ensure the desired delay period of three to six days. Three additional participants were excluded due to procedural errors (two participants received access to and completed part 2 early and one participant completed part 1 twice). Overall, 220 participants (58 male; $M_{age} = 20.51$, $SD = 2.63$, one not reporting) completed both parts of the experiment: 55 in the first wave and 165 in the second wave. Adding time of data collection as a factor to the analysis reveals neither a significant main effect, $F(1, 216) = 0.15$, $p = .703$, $\eta_p^2 < .01$, nor a significant interaction with warning condition, $F(1, 216) = 0.51$, $p = .475$, $\eta_p^2 < .01$, repetition, $F(1, 216) = 1.03$, $p = .312$, $\eta_p^2 = .01$, or three-way interaction, $F(1, 216) = 2.05$, $p = .154$, $\eta_p^2 = .01$. Moreover, analyzing both waves of data collection separately reveals the same pattern of significant main and interaction effects for each wave. Hence, we collapsed analysis across the two waves, resulting in 113 participants in the no warning condition and 107 participants in the pre-exposure warning condition.

Materials. We selected trivia claims from a larger set of previously normed claims (Jalbert, Newman, & Schwarz, 2019). The trivia claims covered a variety of topics (sports, geography, food, animals, and science) and were selected to be ambiguous—only those rated as true between 35% and 65% of the time were included. Examples of true claims are "Walruses use their tusks primarily for mating" and "Kava is a beverage made from the root of the pepper plant." Examples of false claims are "The mouth of a sea urchin is on its top" and "Biking is the first event in a triathlon." During norming, false claims had been created by taking a true claim (e.g., "The mouth of a sea urchin is on its *bottom*"; "*Swimming* is the first event in a triathlon") and altering one word to an incorrect but plausible sounding alternative. We only chose either the true or false version of a claim so that participants would never see both.

Participants saw 36 trivia claims during the initial exposure. In the test phase, participants saw the same 36 claims as well as 36 new claims, for a total of 72 claims. Each claim was presented for 5 sin random order during each session. Half of the trivia claims were true and half were false, both during exposure and testing phases.

The trivia claims were counterbalanced such that half of the participants saw one set of 36 claims repeated, and half of the participants saw the other set of 36 claims repeated. Based on the norming data, both true and false claims were rated as true approximately the same proportion of the time (both $M = 0.52$, $SD = 0.08$).

Procedure. The procedures for all experiments reported in this paper were approved by the University of Southern California's Institutional Review Board (IRB). When participants signed up, they agreed to complete both parts of a two-part online survey.

Exposure phase. Immediately after completing a consent form, participants were told that they would see a series of trivia claims for approximately 3 min. In the pre-exposure warning condition, participants were additionally given a warning "half the statements are true, and half the statements are false." Participants in the no warning condition were not told this. All participants were then told that the trivia statements would be presented automatically and that there was no need to press any buttons. They were asked to read the trivia statements carefully as they were presented, but to not do anything else. The claims were then presented.

After viewing the claims, participants were asked a few general questions to provide a rationale for presenting the claims, including "How many statements do you think you read?" and "How many minutes do you think it took you to read the statements?"

Delay phase. After a three-day delay, participants received a link to part 2 of the survey and were given 48 hours to complete it. Thus, the total time between exposure and test was 72 to 144 hours.

Test phase. In the test phase, participants were shown another series of trivia claims. All participants were correctly told that half of the statements were ones that they had seen before and half of the statements were new. None of the participants were given any warning about the truth of the statements. For each claim, all participants answered the question "Is this statement true or false?" on an unnumbered six-point scale from *definitively true* (coded as 6) to *definitely false* (coded as 1).

Demographics. Following the truth ratings, participants completed individual difference measures, unrelated to the present hypotheses. The results of one of these measures, an 18

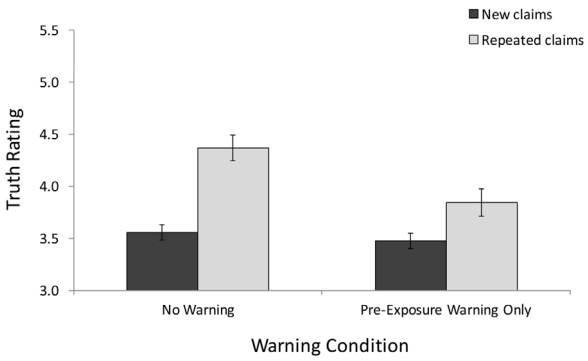

Figure 1. Mean truth ratings across warning conditions for new and repeated claims after a three to six day delay in Experiment 1. Participants either received a warning that half of the claims were false prior to exposure only or no warning. Truth ratings were made on an unnumbered six-point scale from *definitively true* (coded as 6) to *definitely false* (coded as 1). Error bars are 95% confidence intervals.

item Need for Cognition Scale, is reported in Experiment 3 of Newman, Jalbert, Schwarz, and Ly (2020) as a reanalysis of this existing data set. Finally, participants answered demographic questions, including gender and age.

Results

We performed a 2 (warning: before exposure vs. no warning) × 2 (repetition: trivia claim repeated vs. new) mixed model ANOVA. All reported means are estimated marginal means; the raw means show the same pattern and are included in the supplemental materials. Replicating the standard truth effect, repeated claims were rated as more true (M = 4.10, 95% CI [4.02, 4.19]) than new claims (M = 3.51, 95% CI [3.46, 3.57]), mean difference = 0.59, 95% CI [0.51, 0.67], $F(1, 218) = 192.62$, $p < .001$, $\eta_p^2 = .47$, for the main effect. More important, participants who received a warning prior to exposure rated claims as less true (M = 3.66, 95% CI [3.57, 3.74]) than participants who did not receive a warning (M = 3.96, 95% CI [3.88, 4.05]), mean difference = 0.30, 95% CI [0.18, 0.42], $F(1, 218) = 24.58$, $p < .001$, $\eta_p^2 = .10$, for the main effect. These main effects are qualified by an interaction of warning and repetition, $F(1, 218) = 26.88$, $p < .001$, $\eta_p^2 = .11$.

To diagnose this interaction, we computed simple effects using a Bonferroni correction for multiple comparisons. For this analysis and later analyses using a Bonferroni correction, we report a p-value adjusted for these multiple comparisons that can be compared to the standard alpha level of .05. Repeated measures d effect sizes were calculated using Comprehensive Meta-Analysis Software (Version 3.0) from mean differences and standard deviations of the differences, taking into account the correlation between repeated and new claims and corrected for small sample size. This analysis revealed significant truth effects in both conditions (Figure 1). However, the truth effect was considerably larger without a pre-exposure warning, $F(1, 218) = 186.80$, $p < .001$, $\eta_p^2 = .46$, $d = 1.31$, 95% CI [1.02, 1.60], than with a pre-exposure warning, $F(1, 218) = 36.79$, $p < .001$, $\eta_p^2 = .14$, $d = 0.70$, 95% CI [0.48, 0.91]. The difference is driven by participants' truth ratings of repeated claims, which are higher in the absence of a warning (M = 4.37, 95% [4.24, 4.49]) than after a pre-exposure warning (M = 3.84, 95% CI [3.72, 3.97]), mean difference = 0.52, 95% CI [0.35, 0.70], $F(1, 218) = 33.69$, $p < .001$, $\eta_p^2 = .13$. Participants' truth judgments for new statements did not differ across conditions, $F(1, 218) = 2.33$, $p = .128$, $\eta_p^2 = .01$.

The applied importance of these findings is more apparent when one considers how frequently repetition of a claim can shift people's judgments from false to true. We can assess this by analyzing how many claims are rated on the "false" versus "true" side of our six point scale. Recall that the claims were normed to be judged true about half of the time. Consistent with this norming, participants rated new claims to be true about half the time, independent of whether they received no warning (M = 51.18%, SD = 14.50%) or a pre-exposure warning (M = 51.28%, SD = 15.14%). For repeated claims, acceptance as true increased to 70.11% (SD = 16.83%) without a warning but only to 60.23% (SD = 16.77%) with a pre-exposure warning. This reflects a robust illusory truth effect of almost 20 percentage points without a warning that is cut in half with a pre-exposure warning.

In sum, replicating earlier studies, prior exposure to a statement increased its perception as true, even three to six days later. However, alerting participants at the time of exposure that some of the statements they are about to see are false was sufficient to significantly reduce the size of this truth effect. Pre-exposure warnings did not shift participants' ratings of new claims- rather, the reduction in the size of the truth effect was driven by reducing the perceived truth of the repeated claims. A signal detection analysis of response bias (c) parallels these findings across all experiments. These findings are reported in our supplemental materials.

Experiment 2

In the real world, people are rarely alerted to the presence of falsehoods before they are exposed to them. However, it is often possible to alert people to falsehoods after exposure. Empirically, post-exposure warnings are usually less effective in correcting the influence of misinformation than pre-exposure warnings (for a review, see Lewandowsky et al., 2012). Experiment 2 tests whether this holds as well for the influence of repetition. Truth effect experiments often include information prior to test that some or half of the claims are false (i.e., Begg et al., 1992; Brashier et al., 2020; Nadarevic & Erdfelder, 2014; Schwartz, 1982; Silva et al., 2016). In Experiment 2 (preregistered at http://aspredicted.org/blind.php?x=8yz2q5), we tested the influence of warnings at the time of test using three conditions: a pre-exposure and pre-test warning condition, a pre-test warning only condition, and a no warning condition. We expected that a pre-test warning alone would have little effect for two reasons. First, if participants are only warned prior to test, it will be difficult to correct for claims that were already encoded as true during the initial exposure. Second, making a truth rating at test is already drawing attention to truth independent of the warning, so an additional warning that some claims are false will provide little new information. These predictions

are consistent with findings by Nadarevic and Aßfalg (2017) that extensive, but not simple, pre-test warnings may reduce the size of the truth effect.

Method

Design. We used a 3 (warning timing: before exposure and before test, only before test, or no warning) × 2 (repetition: trivia claim repeated vs. new) mixed design, manipulating warning timing between subjects and repetition within subjects.

Participants. We aimed to recruit 300 Mechanical Turk (MTurk) workers, located in the United States, with a HIT approval rate of at least to 95% to complete the experiment using the online survey platform Qualtrics. We again chose to overpower our study and aimed to recruit 100 participants for each of the three between-subjects conditions. Participants were told the experiment would take approximately 30–45 min and were paid $1.20.

A total of 297 participants completed the study. The actual number of participants varies slightly from the posted MTurk HITs because of interactions between Qualtrics and MTurk procedures. However, due to an error in recruitment, 15 participants had taken a past truth effect survey with the same materials, and these participants were excluded, resulting in a remaining sample of 282 participants (127 male; M_{age} = 37.18, SD = 12.13; n = 96 in the pre-exposure and pre-test warning condition, n = 97 in the pre-test warning only condition, n = 89 in the no warning condition).

Procedure. The procedure was similar to the procedure of Experiment 1, except that all parts took place in one experimental session.

Exposure phase. Participants were told that they would see a series of trivia claims for approximately 3 min. Participants in the warning before exposure and before test condition were told that "Half of these trivia statements are true, and half of these trivia statements are false." Participants in the other conditions were not given this warning. After exposure, participants were not asked any general questions, but instead moved immediately to the delay phase.

Delay phase. A twenty-minute delay followed the initial exposure. During that time, participants answered multiple-choice questions about articles unrelated to the trivia claims.

Test phase. The test phase was identical to the test phase of Experiment 1, except for the introduction of a warning prior to test for some participants. All participants were correctly told that half of the statements were ones that they had seen before and half of the statements were new. Following this information, participants in the warning before exposure and before test condition and the warning before test only conditions were given a pre-test warning that half of the statements were true and half of the statements were false. Participants in the no warning condition did not receive this warning.

Demographics. Finally, participants answered demographic questions.

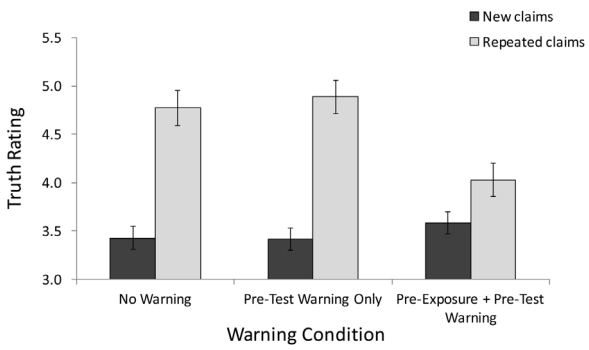

Figure 2. Mean truth ratings across warning conditions for new and repeated claims in Experiment 2. Participants either received a warning that half of the claims were false prior to exposure and test, prior to test only, or did not receive a warning. Truth ratings were made on an unnumbered six-point scale from *definitively true* (coded as 6) to *definitely false* (coded as 1). Error bars are 95% confidence intervals.

Results

We performed a 3 (warning timing: before exposure and before test, only before test, or no warning) × 2 (repetition: trivia claim repeated vs. new) mixed model ANOVA. Replicating the standard truth effect, participants rated repeated claims (M = 4.56, 95% CI [4.46, 4.66]) as more true than new claims (M = 3.48, 95% CI [3.41, 3.54]), mean difference = 1.09, 95% CI [0.97, 1.20], $F(1, 279) = 327.93$, $p < .001$, $\eta_p^2 = .54$, for the main effect of repetition. Moreover, a main effect of warning, $F(2, 279) = 11.92$, $p < .001$, $\eta_p^2 = .08$, reflected that participants who received a warning prior to exposure and prior to test found the statements less true (M = 3.81, 95% CI [3.70, 3.91]), whereas those who only received a warning prior to test (M = 4.15, 95% CI [4.05, 4.26]) reported similar truth judgments as participants who received no warning (M = 4.10, 95% CI [3.99, 4.21]). However, these main effects were qualified by an interaction of warning and repetition, $F(2, 279) = 29.39$, $p < .001$, $\eta_p^2 = .17$.

We followed-up the interaction with simple effects analyses using a Bonferroni correction for multiple comparisons. As shown in Figure 2, and replicating the findings of Experiment 1, the interaction was driven by a change in truth ratings for repeated claims. Repeated claims received significantly higher truth ratings when participants received no warning or a pre-test only warning than when they were warned before exposure (both $p < .001$). The truth ratings for repeated claims in the no warning and pre-test warning only conditions did not differ significantly ($p = 1.00$). Finally, participants' truth ratings of new claims were unaffected by warnings, $F(2, 279) = 2.44$, $p = .089$, $\eta_p^2 = .02$.

A comparison of new and repeated claims showed a significant truth effect in all three conditions. More importantly, this effect was smaller when participants received a pre-exposure and pre-test warning, $F(1, 279) = 18.78$, $p < .001$, $\eta_p^2 = .06$, d = 0.72, 95% CI [0.45, 0.98], than when they received only a pre-test warning, $F(1, 279) = 206.35$, $p < .001$, $\eta_p^2 = .43$, $d = 1.85$, 95% CI [1.37, 2.32], or no warning at all, $F(1, 279) = 159.01$, $p < .001$, $\eta_p^2 = .36$, $d = 1.70$, 95% CI [1.24, 2.15]. In sum, warning participants prior to exposure was critical in reducing the truth

effect by more than half, from $d > 1.70$ to $d = 0.72$. Including only a warning prior to test had no effect.

As noted in Experiment 1, the practical size of these effects is particularly apparent when we consider the proportion of statements rated on the "true" versus "false" side of our 6-point rating scale. Across warning conditions, participants rated similar proportions of new claims as true (no warning: $M = 47.85$, $SD = 18.01\%$; pre-test warning only: $M = 45.93\%$, $SD = 20.75$; pre-exposure and pre-test warning: $M = 52.86\%$, $SD = 17.52\%$). These proportions increased for repeated statements, reflecting sizable illusory truth effects. Without any warning, participants accepted 78.46% ($SD = 19.86\%$) of the statements as true and a pre-test only warning did not affect this proportion (80.41%; $SD = 17.27\%$). In contrast, the combination of a pre-exposure and pretest warning reduced the proportion to 65.09% ($SD = 17.84$), again cutting the size of the illusory truth effect by about half.

Experiment 3

The experimental literature has used two main variations of instructional warnings. In some studies, participants are given information about the specific proportion of true and false claims, usually that "half" of the statements are false (e.g., Begg et al., 1992; Garcia-Marques et al., 2016; Silva et al., 2016). In other studies, participants are merely alerted that "some" statements are false (Brown & Nix, 1996; Brashier et al., 2020; Schwartz, 1982) or that statements could be true or false (e.g., Gigerenzer, 1984; Hasher et al., 1977; Mutter, Lindsey, & Pliske, 1995) without any rate information. In Experiment 3 (preregistered at: https://aspredicted.org/ir6fa.pdf) we explored whether these variations make a difference by comparing both types of warnings ("half" vs. "some") when administered only prior to test or prior to exposure and prior to test.

Method

Design. We used a 2 (warning timing: before test vs. before exposure and before test) × 2 (warning content: "half" vs. "some") × 2 (repetition: trivia claim repeated vs. new) mixed design, manipulating warning timing and warning content between subjects and repetition within subjects.

Participants. As in the previous experiments, we recruited 100 participants for each of four between-subjects conditions from MTurk. Participants were told the experiment would take approximately 30–45 min and were paid $1.20. In total, 405 participants completed the study (154 male; $M_{age} = 37.48$, $SD = 11.55$). One participant was excluded because they reported their age to be less than 18 years old. This left 206 participants in the "some" warning condition ($n = 106$ in the pre-test warning only condition; $n = 100$ in pre-exposure and pre-test warning condition) and 198 participants in the "half" warning condition ($n = 97$ in pre-test warning only; $n = 101$ in pre-exposure and pre-test warning).

Procedure. This study was an exact replication of Experiment 2, with two exceptions. First, we only included two warning conditions, namely a warning prior to exposure and test versus a warning prior to test only. Second, we added a variation in the specific content of the warning, with half of participants receiving the warning that half of the claims true and half of the claims were false and the remaining participants receiving the warning that some of the claims were true and some of the claims were false.

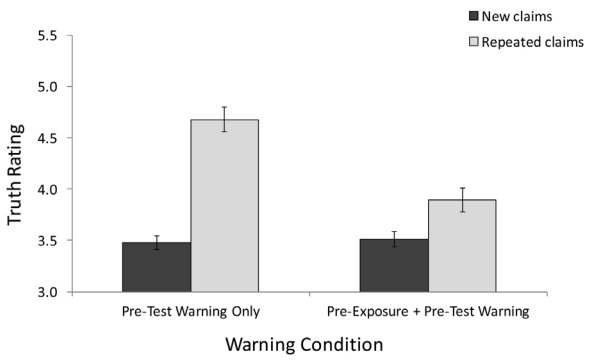

Figure 3. Mean truth ratings for repeated and new claims across warning timing conditions in Experiment 3. Participants received a warning that "some" or "half" of claims were true and "some" or "half" of claims were false. In the pre-exposure and pre-test warning condition, participants received a warning prior to initial exposure to trivia claims and prior to test. In the pre-test warning only condition, participants received a warning prior to test only. Truth ratings were made on an unnumbered six-point scale from *definitively true* (coded as 6) to *definitely false* (coded as 1). Error bars are 95% confidence intervals.

Results

We performed a 2 (warning timing: before test vs. before exposure and before test) x 2 warning content: "half" or "some") x 2 (repetition: trivia claim repeated vs. new) mixed model ANOVA. Whether participants were told that "half" or "some" of the claims were false made no difference, with $F(1, 400) = 0.01$, $p = .933$, $\eta_p^2 < .01$ for the main effect, $F(1, 400) = 0.01$, $p = .935$, $\eta_p^2 < .01$ for the interaction with warning timing, $F(1, 400) = 1.32$, $p = .251$, $\eta_p^2 < .01$ for the interaction with claim repetition, and $F(1, 400) = 0.03$, $p = .611$, $\eta_p^2 < .01$ for the three-way interaction. Thus, this variable is not further discussed.

Replicating the illusory truth effect, participants rated repeated claims ($M = 4.29$, 95% CI [4.20, 4.37]) as significantly more true than new claims ($M = 3.50$, 95% CI [3.45, 3.55]), mean difference = 0.79, 95% CI [0.71, 0.88], $F(1, 400) = 335.25$, $p < .001$, $\eta_p^2 = 0.46$, for the main effect of repetition. As in Experiment 2, participants who were warned prior to exposure and prior to test rated claims as less true ($M = 3.70$, 95% CI [3.63, 3.78]) than participants who were warned prior to test only ($M = 4.08$, 95% CI [4.00, 4.16]), mean difference = 0.38, 95% CI [0.27, 0.48], $F(1, 400) = 46.88$, $p < .001$, $\eta_p^2 = .11$, for the main effect of warning time. These main effects were again qualified by a significant interaction between warning timing and claim repetition, $F(1, 400) = 88.99$, $p < .001$, $\eta_p^2 = .18$.

As shown in Figure 3, simple effects analyses using a Bonferroni correction for multiple comparisons revealed a significant truth effect when a warning was presented only prior to test, $F(1, 400) = 39.24$, $p < .001$, $\eta_p^2 = .09$, $d = 1.47$, 95% CI [1.23, 1.70], and when a warning was presented prior to exposure and repeated prior to test, $F(1, 400) = 386.38$, $p < .001$, $\eta_p^2 = .49$, $d = 0.64$, 95% CI [0.45, 0.84]. As in Experiment 2, the truth effect was significantly smaller in the latter condition. This difference

was again driven by how participants rated repeated claims: these claims were judged as less true when participants received a warning prior to exposure and prior to test ($M = 3.90$, 95% [3.78, 4.01]) than when they received a warning only prior to test ($M = 4.68$, 95% [4.56, 4.80]), mean difference = 0.78, 95% CI [0.62, 0.95], $F(1, 400) = 85.94$, $p < .001$, $\eta_p^2 = .18$. Participants' ratings of new claims were unaffected by the warnings, $F(1, 400) = 0.42$, $p = .517$, $\eta_p^2 < .01$.

To assess the practical size of these effects, we again consider the proportion of claims rated on the "true" versus "false" side of the rating scale. Consistent with the norming of the ambiguous claims, participants accepted new claims as true about half the time across conditions (pre-test warning only: $M = 48.91\%$, $SD = 18.20\%$; pre-exposure and pre-test warning: $M = 50.25\%$, $SD = 15.48\%$). For repeated claims, acceptance as true increased to 75.86% ($SD = 21.19\%$) in the pre-test warning only condition, while adding a pre-exposure warning reduced the proportion to 60.97% (SD = 18.14%).

In sum, these findings are consistent with Experiments 1 and 2. Warning participants about the presence of falsehoods prior to exposure reduced the size of the truth effect by more than half, even when all participants were given a warning prior to test. There was no difference between a warning that "some" of the claims were false and a warning that "half" of claims were false.

Effect Size Analyses

Figure 4 shows a forest plot of all effect size estimates for each warning timing and warning type condition. Analysis was performed using Comprehensive Meta-Analysis Software (Version 3.0). Due to the small number of studies, tau-squared was pooled across studies, following recommendations by Borenstein, Hedges, Higgins, & Rothstein (2009). A random effects model was used and effect sizes were fixed across subgroups. Effect sizes were corrected for small sample biases (Borenstein et al., 2009). The effects were grouped into three subgroups: conditions where no warnings were given, conditions where warnings were only given at the time of test, and conditions where warnings were given at the time of exposure (both with and without repetition of the warning at the time of test). Each effect fit into exactly one of these three subgroups, so no effects were double-counted.

The total effect size across all conditions was $d = 1.00$, 95% CI [0.90, 1.10]. There was evidence of a significant difference among warning conditions, $Q(2) = 58.61$, $p < .001$. Follow-up analyses using a Bonferroni correction for multiple comparisons confirmed that the truth effect was significantly smaller in the pre-exposure warning condition, $d = 0.68$, 95% CI [0.55, 0.81], than in both the pre-test warning only condition, $d = 1.55$, 95% CI [1.33, 1.76], and the no warning condition, $d = 1.42$, 95% CI [1.18, 1.67]; both $Q(1) > 28$, $p < .001$. The pre-test warning only and no warning conditions were not significantly different, $Q(1) = 0.27$, $p = 1.000$.

Additional Analyses

What accounts for the robust observation that pre-exposure warnings reliably reduce the size of the truth effect? One possibility is that pre-exposure warnings increase participants' ability to accurately discriminate between true and false repeated claims at test. To test this possibility, we performed a signal detection analysis (Stanislaw & Todorov, 1999) using discrimination (***d'***) to investigate whether participants who received pre-exposure warnings were more accurate at discriminating between true and false claims. Following earlier research (e.g., Begg et al., 1992; Garcia-Marques et al., 2016; Mitchell et al., 2006), we converted the unnumbered six-point scale used in the experiments (ranging from *definitely true*, coded as 6, to *definitely false*, coded as 1) to a binary measure, with values from 1 to 3 treated as "false" and values from 4 to 6 treated as "true." Across all three studies, there was no consistent influence of warning timing on participants' ability to discriminate between true and false claims. In Experiments 1 and 2, there was no significant

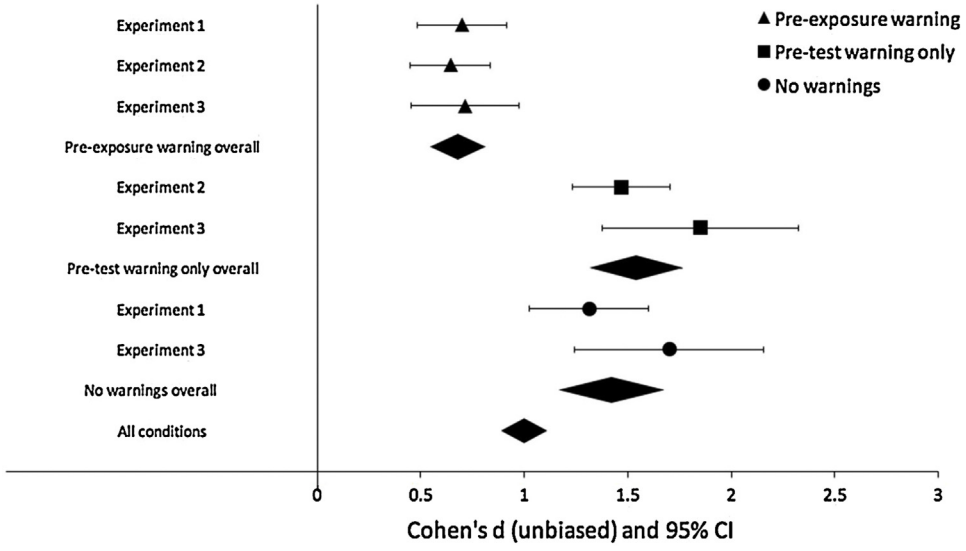

Figure 4. Effect sizes (*d* unbiased) for the truth effect across all experiments and warning timing conditions. These effects represent 95% confidence intervals.

main effect of warning timing; Experiment 1: $F(1, 218) = 2.64$, $p = .105$, $\eta_p^2 = .01$; Experiment 2: $F(2, 279) = 0.26$, $p = .774$, $\eta_p^2 < .01$. Additionally, neither experiment had a significant interaction of warning timing and repetition, both Fs < 0.99, ps > .373. In Experiment 3, no main effect of warning timing emerged, $F(1, 400) = 0.04$, $p = .850$, $\eta_p^2 < .01$, but one significant interaction with warnings did: the three way interaction of repetition, warning timing, and warning content, $F(1, 400) = 5.60$, $p = .018$, $\eta_p^2 = .01$. However, simple effects analysis using Bonferroni correction for multiple comparisons revealed that pre-exposure warnings did not significantly increase overall accuracy in either the condition using a "some" warning or the condition using a "half" warning (both Fs < 0.84, ps > .512), nor did they significantly increase accuracy for any specific claim type (new or repeated) within those conditions. In fact, the only single significant effect when looking at new and repeated claims alone was in the opposite direction, with pre-exposure warnings decreasing accuracy for repeated claims in the "half" warning condition relative to a pre-test warning only condition (pre-test warning only: $M = 0.22$, 95% CI [0.13, 0.33]; pre-test and pre-exposure warning: $M = 0.07$, 95% CI [-0.03, 0.17]), raw mean difference = -0.15, 95% CI [-0.29, 0.00], $F(1, 400) = 3.96$, $p = .047$, $\eta_p^2 = .01$ (all Fs < 1.91, ps > .170 in other conditions). In short, we found no evidence that pre-exposure warnings reduce the size of the truth effect by improving participant's ability to accurately discriminate between true and false claims.

Another possibility is that warnings at exposure allow participants to catch inconsistencies in false claims that they may otherwise not notice, protecting participants from accepting the claim at test. This would be consistent with the finding of Brashier et al. (2020) that asking participants to fact-check information at exposure reduces the size of the truth effect for false claims, provided they have the relevant knowledge to draw on. If so, only factually false claims should show a smaller truth effect in the pre-exposure warning conditions. This prediction entails a three-way interaction of actual truth value of the claim, repetition, and warning.

To test this possibility, we ran additional analyses with the actual truth value of the claims added as a within-subjects factor. Despite norming that aimed to equalize the perceived truth of true and false claims, we found a main effect of actual truth value in each experiment. Participants consistently rated true claims as more true than false claims, Experiment 1: $F(1, 218) = 72.13$, $p < .001$, $\eta_p^2 = .25$; Experiment 2: $F(1, 279) = 51.83$, $p < .001$, $\eta_p^2 = .16$; Experiment 3: $F(1, 400) = 51.65$, $p < .001$, $\eta_p^2 = .11$, although this effect was relatively small (raw mean difference on a six point scale: Experiment 1: 0.26, 95% CI [0.20, 0.32]; Experiment 2: 0.21, 95% CI [0.15, 0.27]; Experiment 3: 0.18, 95% CI [0.13, 0.23]). More importantly, the three-way interaction of truth value, warning timing, and repetition was $F < 1$ in each of the experiments, Experiment 1: $F(1, 400) = 0.09$, $p = .771$, $\eta_p^2 < .01$; Experiment 2: $F(2, 279) = 0.77$, $p = .464$, $\eta_p^2 < .01$; Experiment 3: $F(1, 218) = 0.15$, $p = .703$, $\eta_p^2 < .01$. The consistent lack of three-way interactions indicates that the impact of warning timing on the size of the truth effect was not moderated by the actual truth value of the claims. See the supplemental materials for a detailed report of this analysis.

General Discussion

Across three experiments, a remarkably consistent pattern emerged: pre-exposure warnings that alerted participants that some of the claims they were about to see would be false significantly reduced the size of the truth effect compared to all other conditions. In contrast, warnings given only prior to test exerted no influence and failed to reduce the truth effect compared to a condition without any warning—likely because asking participants to rate truth at test already draws attention to the truth value of the claim. The protective effect of pre-exposure warnings was observed both when people were told that "half" or "some" of the statements were false. Given that warnings at exposure are a common feature of truth effect studies, this robust finding implies that the bulk of this literature underestimates the influence of repetition under conditions in which falsehoods do not come with a warning label.

Based on a meta-analysis of seventy effect sizes, Dechêne et al. (2010) reported an average between-items truth effect size of $d = 0.50$, 95% CI [0.43, 0.57]. The confidence interval of their estimate overlaps with the confidence interval of the effect sizes of all conditions that included a pre-exposure warning in the present experiments, $d = 0.68$, 95% CI [0.55, 0.81]. This is consistent with the observation that most prior truth effect studies warned participants about truth at initial exposure, either by asking participants to rate the truth of each claim (51 out of 70 effect sizes included in Dechêne et al., 2010) or by explicitly informing them that some of the claims they will see are false. However, the present three experiments converge on much higher effect size estimates when participants received no warning at all, $d = 1.55$, 95% CI [1.33, 1.76], or were only warned prior to test, $d = 1.42$, 95% CI [1.18, 1.67]. These effect sizes are two to three times the size otherwise reported, suggesting that the experimental procedures commonly used in truth effect studies are likely to underestimate the impact of message repetition on later judgments of truth under natural conditions.

While results strongly support this conclusion of applied relevance, they also allow us to consider possible underlying mechanisms. One is that alerting people prior to exposure that not all of the statements are true makes them more careful at test—a criterion shift. If so, one would expect participants to be less likely to rate any claim as true at test. Our results make this unlikely: pre-exposure warnings influenced participants' responses to repeated claims, but not new claims. If warnings made people more careful overall, we should have seen a reduction in true responses for both new and repeated claims. Instead, only repeated claims were rated less true after pre-exposure warnings. Signal detection analyses with response bias (c) align with these findings. Moreover, an identical warning given only immediately before test had no influence. Another possibility is that pre-exposure warnings allow participants to identify false claims during the exposure phase, which they would otherwise not have noticed. When these false claims are seen again at test, they may be more likely to be correctly identified as false. If so, pre-exposure warnings should protect participants from believing factually false claims, but should do little to reduce belief in factually true claims. Our results do not support this

account either—the influence of warnings on the size of the truth effect was not modified by the actual truth value of claims. Additionally, warnings had no influence on participant's ability to accurately discriminate between factually true and factually false claims at test.

An additional possibility is that participants use the recollection that they saw the claim during the exposure phase in evaluating the credibility of the source: when they were warned that some of the presented claims are false the source is less credible than without such a warning. Supporting this possibility, several studies found that recollection can reduce acceptance of a claim under such conditions (e.g., Begg et al., 1992; Brown & Nix, 1996; Unkelbach & Stahl, 2009), although this is not always the case (Henkel & Mattson, 2011). Because recollection becomes less likely as time passes, this account predicts that a pre-exposure warning should be more effective when the delay between exposure and test is short. In contrast, we observed a similar influence of pre-exposure warnings for delays of 20 min (Experiments 2 and 3) and delays of 3 to 5 days (Experiment 1), which renders this account less likely.

Instead, we propose that variables that alert participants to the potential presence of falsehoods at the time of exposure are effective because they disrupt the tacit assumptions underlying communication in daily life. As observed in many studies of conversational pragmatics, people generally assume that communicated information is truthful, unless contextual cues or other knowledge indicate that the communicator may not be cooperative (Grice, 1975; Sperber & Wilson, 1986; for reviews, see Schwarz, 1994, 1996). When contextual cues elicit distrust, recipients consider how things might differ from what is claimed (for a review, see Mayo, 2015), resulting in a higher accessibility of incongruent associations (Schul, Mayo, & Burnstein, 2004) and more counterfactuals and alternative accounts (Kleiman, Sher, Elster, & Mayo, 2015, & Mayo, 2015; Schul, 1993). For example, people who are warned that information may be false prior to reading the false claim "Biking is the first event in a triathlon" may be more likely to think about whether people actually start a triathlon by running or swimming. When later assessing the claim's truth, these contradictory possibilities would more easily come to mind as plausible alternatives, reducing the impact of repetition on truth.

This possibility is consistent with Unkelbach and Rom's (2017) observation that bringing associations to mind that are incongruent with a given claim attenuates the otherwise observed truth effect, which may reflect an influence of the accessible declarative information (Brashier & Marsh, 2020; Unkelbach et al., 2019) or the reduced fluency experienced when processing contradictory information (Winkielman et al., 2012). Paralleling these observations, distrust at the time of encoding has also been found to protect people against the continued influence of misinformation after it is retracted (Fein et al., 1997; Lewandowsky et al., 2005, 2012). From this perspective, any manipulation that fosters distrustful elaboration at encoding may counteract the influence of repetition-based familiarity.

Finally, a methodological point is worth noting. How people process information they encounter in an experiment depends on what they consider their task to be. As discussed in the introduction, most illusory truth experiments draw attention to the truth value of the claims at an early stage, either through instructional warnings or early requests to evaluate the truth of the claim. As seen, such tasks can decrease the size of illusory truth effects. Other processing tasks, like rating how interesting or easy to understand a statement is, are likely to draw attention away from truth. This may increase the size of illusory truth effects when they prompt detailed encoding of the statement without attention to its veracity, thus facilitating more fluent processing at test. Examining the impact of different initial exposure conditions on the size of the illusory truth effect provides a fruitful area for future research with important theoretical and applied implications.

Applied Implications

The present findings indicate that the experimental literature on illusory truth effects has most likely underestimated the truth-creating power of repetition in everyday life. This is the case because widely used experimental procedures alert participants that some of the claims are false. In our experiments, the impact of repetition was two to three times larger when these pre-exposure warnings were removed. This is consistent with the finding that warnings can provide protection against the acceptance of misinformation in different domains (Lewandowsky et al., 2012). Given that false information rarely comes with warning labels in the real world, the latter effect size estimates are probably closer to what may be observed under natural conditions, provided that the communicator is not perceived as untrustworthy. Additionally, truth effect experiments are typically limited to a single repetition, as were the present studies. In contrast, social media and the 24-hour cable news cycle ensure a much larger number of repetitions, which may further enhance the size of truth effects (cf. Hasher et al., 1977).

On a more optimistic note, our findings suggest that simple reminders that not all information is true can protect people from believing false information when they encounter it again. Including such reminders in contexts where false information is likely to be present, such as at the top of social media feeds or articles that have not been fact checked, may provide an opportunity to curb later belief in false information in contexts where more extensive interventions may be impossible.

As this discussion indicates, much remains to be learned about the applied implications of the robust illusory truth effects documented in dozens of laboratory studies since Hasher et al. (1977)'s pioneering work. Hopefully, future research that more closely mirrors how people encounter information in daily life will fill these gaps, providing a fuller understanding of the role of repetition in the creation and maintenance of beliefs.

Author Contributions

All three authors conceived and designed the studies. Madeline Jalbert collected, analyzed, and interpreted the data with assistance by Eryn Newman. All three authors wrote the manuscript and approved the final version for submission.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Begg, I., 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*, 446–458.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Brashier, N. M., Eliseev, E. D., & Marsh, E. J. (2020). An initial accuracy focus prevents illusory truth. *Cognition*, *194*, 104054.

Brashier, N. M., & Marsh, E. J. (2020). Judging truth. *Annual Review of Psychology*, *71*, 499–515.

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1088–1100.

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*, 238–257.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Fein, S., McCloskey, A. L., & Tomlinson, T. M. (1997). Can the jury disregard that information? The use of suspicion to reduce the prejudicial effects of pretrial publicity and inadmissible testimony. *Personality and Social Psychology Bulletin*, *23*, 1215–1226.

Garcia-Marques, T., Silva, R. R., & Mello, J. (2016). Judging the truth-value of a statement in and out of a deep processing context. *Social Cognition*, *34*, 40–54.

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology*, *97*, 185–195.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics, Vol.3: Speech acts* (pp. 41–58). New York: Academic Press.

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112.

Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, *19*, 212–225.

Henkel, Linda, & Mattson, Mark. (2011). Reading is believing: the truth effect and source credibility. *Consciousness and Cognition*, *20*(4), 1705–1721. https://doi.org/10.1016/j.concog.2011.08.018

Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 21–38.

Jalbert, M., Newman, E., & Schwarz, N. (2019). Trivia claim norming: Methods report and data. *ResearchGate.*, https://doi.org/10.6084/m9.figshare.9975602

Johnson-Laird, P. N. (2012). Mental models and consistency. In B. Gawronski, & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 225–243). New York: Guilford Press.

Kleiman, T., Sher, N., Elster, A., & Mayo, R. (2015). Accessibility is a matter of trust: Dispositional and contextual distrust blocks accessibility effects. *Cognition*, *142*, 333–344.

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*, 1093–1096.

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

Lewandowsky, S., Stritzke, W. G., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation: The Iraq War 2003. *Psychological Science*, *16*, 190–195.

Mayo, R. (2015). Cognition is a matter of trust: Distrust tunes cognitive processes. *European Review of Social Psychology*, *26*, 283–327.

Mitchell, J. P., Sullivan, A. L., Schacter, D. L., & Budson, A. E. (2006). Misattribution errors in Alzheimer's disease: The illusory truth effect. *Neuropsychology*, *20*, 185–192.

Mutter, S. A., Lindsey, S. E., & Pliske, R. M. (1995). Aging and credibility judgment. *Aging and Cognition*, *2*, 89–107.

Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: warnings reduce the repetition-based truth effect. *Psychological Research*, *81*, 814–826.

Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, *23*, 74–84.

Newman, E. J., Jalbert, M. C., Schwarz, N., & Ly, D. P. (2020). Truthiness, the illusory truth effect, and the role of Need for Cognition. *Consciousness and Cognition*, *78*, 102866.

Newman, E. J., Sanson, M., Miller, E. K., Quigley-McBride, A., Foster, J. L., Bernstein, D. M., et al. (2014). People with easier to pronounce names promote truthiness of claims. *PLOSone*, *9* https://doi.org/10.1371/journal.pone.0088671

Newman, E. J., & Schwarz, N. (2018). Good sound, good research: How audio quality influences perceptions of the research and researcher. *Science Communication*, *40*, 246–257.

Parks, C. M., & Toth, J. P. (2006). Fluency, familiarity, aging, and the illusion of truth. *Aging, Neuropsychology, and Cognition*, *13*, 225–253.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338–342.

Schul, Y. (1993). When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology*, *29*, 42–62.

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*, 668.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology*, *26*, 123–162.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods and the logic of conversation*. Hillsdale, NJ: Erlbaum.

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review*, *2*, 87–99.

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & E. R. Smith (Eds.), *The mind in context* (pp. 105–125). New York: Guilford Press.

Schwarz, N. (2015). Metacognition. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA Handbook of Personality and Social Psychology: Attitudes and Social Cognition* (pp. 203–229). Washington, DC: APA.

Schwarz, N. (2018). Of fluency, beauty, and truth: Inferences from metacognitive experiences. In J. Proust, & M. Fortier (Eds.), *Metacognitive diversity. An interdisciplinary approach* (pp. 25–46). New York: Oxford University Press.

Schwartz, M. (1982). Repetition and rated truth value of statements. *American Journal of Psychology*, *95*, 393–407.

Silva, R. R., Garcia-Marques, T., & Mello, J. (2016). The differential effects of fluency due to repetition and fluency due to color contrast on judgments of truth. *Psychological Research*, *80*, 821–837.

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*, 713–724.

Sperber, D., & Wilson, D. (1986). . *Relevance: Communication and Cognition* (Vol. 142) Cambridge, MA: Harvard University Press.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments & Computers*, *31*, 137–149.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, *28*, 247–253.

Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, *160*, 110–126.

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38.

Winkielman, P., Huber, D. E., Kavanagh, L., & Schwarz, N. (2012). Fluency of consistency: When thoughts fit nicely and flow smoothly. In B. Gawronski, & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 89–111). New York: Guilford Press.

Received 25 February 2020;
received in revised form 30 June 2020;
accepted 19 August 2020
Available online 24 November 2020