This is an unpublished working paper. Post date: 10/28/21

Does wording matter? Examining the effect of phrasing on memory for negated political fact-checks

Raunak M. Pillai

Sarah Brown-Schmidt

&

Lisa K. Fazio

Vanderbilt University

Author Note

Raunak M. Pillai, Sarah Brown-Schmidt, and Lisa K. Fazio, Department of Psychology and Human Development, Vanderbilt University.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1937963. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Raunak Pillai, Department of Psychology and Human Development, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203. Email: raunak.m.pillai@vanderbilt.edu

Abstract

After encountering negated messages, people may remember the core claim while forgetting the negative evaluation. These memory errors are of particular concern for fact checks on social media, which often use brief affirmations or negations to help the public learn the truth behind questionable claims. Across three experiments, we examined whether these memory errors could be minimized by placing evaluations before the entire claim is stated (e.g., “No, X did not do Y, as A claims”), rather than after (e.g., “A claims X did Y. No, this is false”). Participants remembered whether fact-checked political claims were affirmed or negated immediately (Experiment 1) and one week later (Experiment 2). While participants began to forget these fact-checks after three weeks, this forgetting was similar for before- and after-claim evaluations, contrary to our predictions (Experiment 3). These results suggest that there are multiple, equally memorable formats for communicating affirmations and negations.

General Audience Summary

In recent years, there has been a growing number of new fact-checking projects, with each broadly seeking to verify and communicate the truth of questionable claims circulating amongst the public. These efforts often include brief social media posts that introduce a claim and then affirm or negate it (e.g., “A viral story claims a Florida school ditched Common Core and then soared to No. 1. False.”). A critical challenge is ensuring that people who see these fact checks remember them over time. We hypothesized that one way to improve memory for these fact checks would be to rephrase them such that the affirmation or negation comes before the claim is fully stated (e.g., “No, this Florida school did not ditch Common Core and then soar to No. 1, as a viral story claims.”), rather than after the claim. Across three experiments, we had participants read a series of fact-checks from PolitiFact about true and false claims regarding contemporary American politics, and then rate the truth of a series of claims. Overall, these fact checks were effective. Participants were more accurate at rating the truth of claims immediately after reading related fact checks, and they were able to remember whether each claim was true or false over the course of one week. After three weeks, however, participants did begin to forget whether claims were true or false. Most importantly though, participants began to forget both types of fact checks at similar rates, contrary to our predictions. Overall, this research contributes to a growing literature suggesting that fact checks can improve the accuracy of people’s factual beliefs, and suggests that there are multiple, equally effective ways of communicating fact checks in the context of social media.

Does wording matter? Examining the effect of phrasing on memory for negated political fact checks

Journalists are often tasked with not only reporting accurate information, but also verifying the accuracy of questionable claims being publicly circulated. These fact-checking endeavors have increased in recent years amid concerns about the rapid circulation of false information, with at least 30 new fact-checking projects launching each year from 2014 to 2020 (Stencel & Luther, 2021). An emerging consensus indicates that these fact checks can be effective at improving readers' knowledge and reducing their beliefs in false claims (see Walter, Cohen, Holbert, & Morag, 2020; Walter & Murphy, 2018 for meta-analyses). However, the effectiveness of different fact-checking formats is still unclear. In particular, there is strong support for the usefulness of in-depth fact checking articles that present both the false claim and why it is incorrect (e.g., Fridkin, Kenney, & Wintersieck, 2015; Nyhan, Porter, Reifler, & Wood, 2020; Weeks, 2015). However, many readers never reach the full article and are instead presented with only the headline or a very brief description on social media. A key challenge for this growing line of work involves how to present these short-form debunks.

One concern with social-media-based fact checks is that they may be too short to affect beliefs. Within the span of a 280-characters limit found on Twitter, for instance, a fact check may only be able to describe a claim and label it as false, without providing additional details. A related concern is that, in this format, much of the space goes to restating the claim in question. Repetition of claims often increases belief (see Unkelbach, Koch, Silva, & Garcia-Marques, 2019 for a review), and so there is a concern that that these fact checks may unintentionally increase belief in the falsehood.

Two recent investigations into the effectiveness of these short-form fact checks have found no evidence of the predicted harms (Ecker, Lewandowsky, & Chadwick, 2020; Ecker, O'Reilly, Reid, & Chang, 2020). In these studies, participants were exposed to brief 140-character fact checks, like those posted to Twitter. The simplest versions simply stated the claim and added a color-coded "true" or "false" label below. Contrary to predictions, the fact checks were effective at increasing the accuracy of people's beliefs relative to both pre-exposure ratings and ratings provided by a separate, unexposed control group.

While these simple fact checks were still effective after delays of about 1 day or 1 week, they were much less effective after the week-long delay, likely because people were beginning to forget which claims were identified as true or false. In sum, the key concern with short-form, social-media-based fact checks is not that they are ineffective (Ecker, O'Reilly, et al., 2020), or that they may inadvertently increase belief in falsehoods (see Swire-Thompson, DeGutis, & Lazer, 2020 for a review). Rather, the critical challenge is ensuring that their content is retained in memory over longer time periods.

Processing and retaining negated ideas

Current psychological accounts of how people process negated messages can help inform why the efficacy of these simple fact checks may fade over time. In particular, one prominent account of negation processing, the "schema-plus-tag model," suggests that, when exposed to a negated message (e.g., "they do not feel cold"), people first form a representation of the core affirmative supposition ("they feel cold") to which a negative "tag" is then affixed (Clark & Chase, 1972; Mayo, Schul, & Burnstein, 2004; see discussion of models of negation processing in Mayo, 2019).

Importantly, a prediction of the schema-plus-tag model is that, over time, people may begin to lose the association between the affirmative schema and its negative tag in memory (Mayo et al., 2004). In the context of fact-checking and belief, these dissociation errors would result in people being less likely to reject claims as false over time, as the association between the claim and its "false" label becomes less accessible.

Variation in negation formats

If dissociation errors are the reason why short, social-media-based fact checks become less effective over time, then one remedy is to construct fact checks that elicit more durable schema-tag associations in memory. One way of achieving this may be to vary how the negated sentence is worded. For example, past studies (Ecker, Lewandowsky, et al., 2020; Ecker, O'Reilly, et al., 2020) have focused on simple fact checks wherein a claim is stated in its entirety and then a label is placed after the claim (e.g., "In 2016, US violent crimes increased 5.3% compared to the same time period in 2015. TRUE"). However, on social-media, fact checks also often present the negative operator in conjunction with and/or before the entirety of the claim is stated, as in the tweet "No, the UN is not planning to implant the world with biometric IDs" (PolitiFact, 2020).

There are several reasons to predict that fact checks formatted with a negation before the claim is fully stated (e.g., "X did not do Y, as A claims") may be remembered better than fact checks which place the negation after the claim (e.g., "A claims X did Y. No, this is false"). Research on the "continued influence effect" suggests that, even in the face of subsequent retractions, exposure to false information can have strong impacts on people's reasoning, (see Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012 for a review). Accordingly, labelling a claim as "false" after the claim is stated in its entirety may be ineffective. Instead, placing the

"false" label before the claim may serve as a form of warning about the presence of falsehoods -- a successful strategy for minimizing the misleading effects of falsehoods in various research paradigms (e.g., Ecker, Lewandowsky, & Tang, 2010; Jalbert, Newman, & Schwarz, 2020; Schul, 1993). These warnings may serve as a signal for distrust, triggering skeptical encoding processes which bring attention to the opposite of what is being presented (Schul, Mayo, & Burnstein, 2004). In the case of fact checks, if the "evaluation before claim" format elicits this kind of distrustful processing, it may result in a stronger association between the claim and its label as "false" in memory. Finally, placing the negative operator before the claim minimizes the length of text during which the false claim is presented without qualification. In this format, the negative operator is typically in the middle of the claim, accompanying the main verb ("X did *not* do Y, as A claims."), whereas placing the evaluation after the claim places the negative operator after the claim is stated in its entirety ("A claims X did Y. This is *not true*."). This temporal proximity between the negation and the main verb may result in a stronger integration between the representation of the claim and its negative tag, improving retention over time.

However, there is also some evidence to suggest that placing negations after the claim may be more effective. Recent studies suggest that providing simple true-false labels immediately after participants finish reading a given news headline results in more accurate truth ratings one week later relative to seeing the same labels prior to or concurrently with the headlines (Brashier, Pennycook, Berinsky, & Rand, 2021). One interpretation is that placing the label after the headline lets people generate an initial prediction regarding the truth status of the headline. The label then serves as feedback, highlighting any discrepancies between the predicted and the actual truth status of the headline. These moments of surprise may then improve memory for the association between the claim and its label (e.g., Fazio & Marsh, 2009).

Thus, corrections may be more effective if they occur after readers have processed the initial claim.

Current Studies

The current studies are designed to examine how people remember the content of short-form, social-media-based fact checks written in two formats: with a negation/affirmation before the claim or with that evaluation after the claim. All three studies used actual fact-checking tweets from PolitiFact as the stimuli. Experiment 1 investigates the efficacy of these short fact checks in improving the accuracy of people's pre-existing beliefs, and whether there were any differences by format in this initial process. Participants read a series of fact checks that evaluated (i.e., affirmed or negated) a set of true or false political claims with the evaluation placed "before" or "after" the claim. Then, participants rated the truth of a full set of claims, including both claims that were initially fact-checked and claims they had not previously encountered.

Experiments 2 and 3 examine our key question: are fact checks better retained over time when they are formatted with an evaluation before the claim? In Experiment 2, participants read a series of fact-checking tweets, and rated the truth of the corresponding claims immediately and/or after a 1-week delay. Experiment 3 was a direct replication of Experiment 2 with a 3-week delay. A key prediction of the schema-plus-tag model is that, over time, memory for the truth of each fact-checking claim will fade. Critically, we hypothesized that this forgetting would occur less often for fact checks with an evaluation before the claim, as they provide opportunities for a stronger association between the claim and its label. By contrast, there is also reason to predict that fact checks with an evaluation after the claim would be better retained, if they afford a greater chance for learning driven by strong prediction errors. Data, pre-registrations, stimuli,

and supplementary analyses for all experiments are available at the project's OSF site: https://osf.io/ndru4/?view_only=ca9c1209c94f449ea5ca029babf1fadb

Experiment 1

The goal of Experiment 1 was to examine the effects of exposure to simple, evaluative fact checks on belief in the fact-checked claim. Participants encountered a series of fact checks, then rated the truth of both fact-checked and novel claims. Past work suggests that these fact checks would be effective—that is, exposure to fact checks would make participants rate true claims as more true and false claims as less true (Ecker, O'Reilly, et al., 2020). However, unlike in past work, exposure to fact checks in this experiment was manipulated within-subjects; participants rated a mixed list of fact-checked and novel claims.

One concern about short evaluative fact checks is that they repeat misinformation, which may make falsehoods seem more familiar or easier to process and thus more true (see Swire-Thompson et al., 2020 for a review). If these effects exist, they are more likely to be observed in mixed lists of repeated and unrepeated information. In mixed lists, the ease of processing repeated information is more distinct from the relative baseline established by other nearby items and thus more influential in judgements (Dechêne, Stahl, Hansen, & Wänke, 2009). In this way, the present design poses a stronger test of concerns that fact checks may unintentionally increase belief in the negated false information. Still, in line with recent evidence, we hypothesized that fact checks would be effective, and in fact decrease belief in falsehoods, while increasing belief in truths. Further, we hypothesized that this immediate effect would occur similarly regardless of fact check format (i.e., evaluation before or after claim).

Method

Participants. One hundred adult participants (M_{age} = 36.88, SD = 11.79, one not reporting) were recruited from Amazon's Mechanical Turk Platform (MTurk) to complete the study online through the Finding Five platform (FindingFive Team, 2019) for a payment of $1.81. We restricted our sample to participants from the United States who had at least a 95% approval rating on MTurk and who had not completed similar studies from our lab in the past. The number of participants matched our pre-registered sample size.

Design. The experiment had a 2 (claim veracity: true, false) × 2 (fact check exposure: new claim, initially fact-checked claim) × 2 (fact check format: evaluation before claim, evaluation after claim) × 2 (fact check origin: journalist, researcher) within-subjects design. Note that for "new claims" for which participants were not exposed to fact checks, there was no defined "fact check format" or "fact check origin." For comparison purposes, we arbitrarily pre-assigned half of the new claims to each level of these variables.

Materials. A set of 40 fact-checking tweets that affirmed or negated political claims were identified from the Twitter account of the fact-checking website PolitiFact (https://twitter.com/PolitiFact). Selected tweets were posted between June 2016 and February 2020, and covered a range of contemporary political issues, including platforms of political candidates, economic statistics, and election fraud. Tweets covered regional, national, and international topics, and varied in partisanship (i.e., tweets evaluated claims from Republicans, Democrats, as well as other sources).

Critically, half of the 40 selected tweets affirmed true claims, and half negated false claims, and, for each kind of claim, half of the tweets provided the evaluation before the claim, and half provided the evaluation after the claim (see Figure 1). Further, for each tweet, we generated a second "researcher" generated version which fact-checked the same claim, but with

the opposite fact-checking format (before vs. after). For instance, when a fact check's original, journalist-written format had a negation before the claim (e.g., "No, this Florida school did not ditch Common Core and then soar to No. 1, as a viral story claims."), we generated a "researcher" version of the same fact check with a negation after the claim (e.g., "A viral story claims a Florida school ditched Common Core and then soared to No. 1. False."). This process ensures that any effects of fact check format are due to the wording manipulation itself, rather than to other factors (e.g., subject matter, partisanship, emotionality) that may naturally co-vary with how journalists choose to word their tweets. Note that, while participants saw a mix of "before" and "after," "journalist" and "researcher" sourced fact checks, for a given fact check about a given claim, participants only saw one of the two possible tweets.

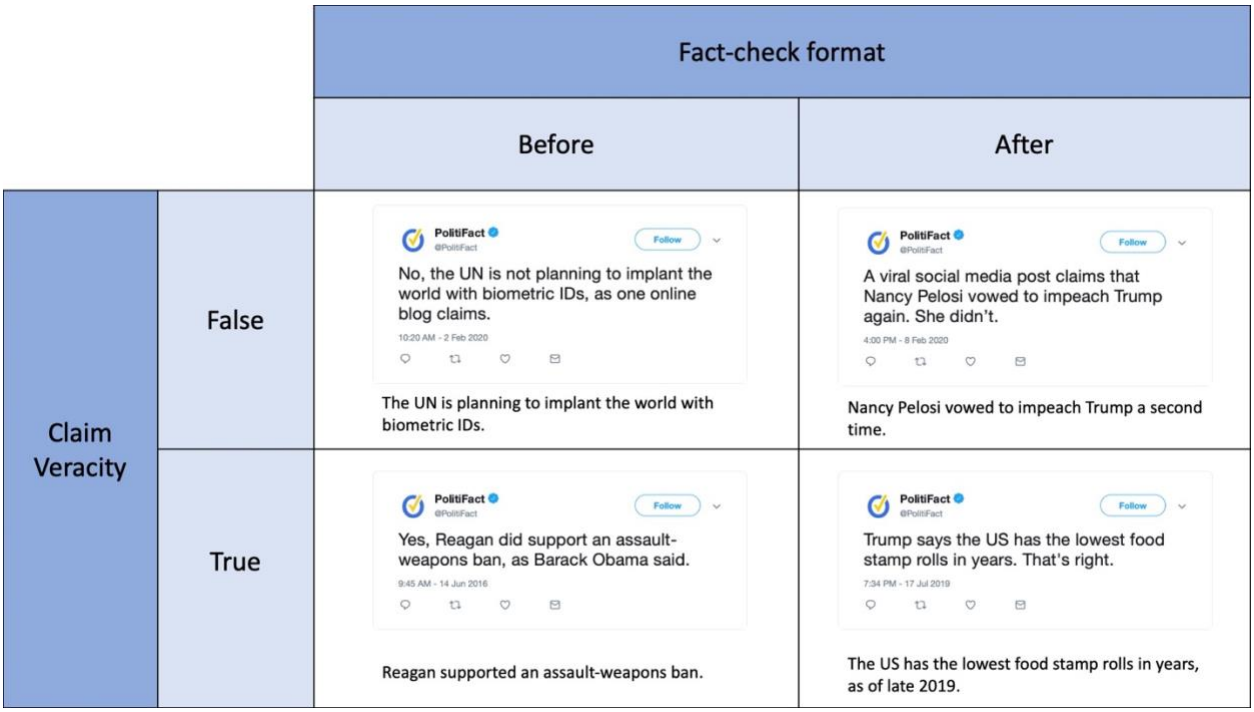

Figure 1. Sample fact-checking tweet screenshots evaluating true (bottom) and false (top) claims, employing an evaluation before (left) and after (right) the claim. The text of claims

shown to participants during the truth-rating phase are shown below each corresponding screenshot.

We then made some slight modifications to the tweets for the purpose of our experiments. First, we ensured all tweets referenced a source of the respective claim (e.g., "…as one Facebook post claims") for uniformity. Second, we removed any extra affirmations or negations of the claim that were originally in the tweet for emphasis (e.g., removing "That's false!" from a tweet in the form "No, X did not do Y. That's false!") so that such tweets could be decisively labelled as having an evaluation "before" the claim. Third, we made some wording edits to increase uniformity (i.e., removing introductory phrases like "NEW:" at the beginning of some tweets) or comprehensibility (i.e., changing "That's Pants on Fire." to "That's false."). Finally, we double checked all tweets for grammar, and added periods at the end of all tweets. As demonstrated in Figure 1, the text of each fact check was formatted to look like a screenshot of a tweet from the PolitiFact Twitter account from the appropriate date and time of the fact check. For uniformity, we removed all URLs and images from these screenshots, and similarly did not include any social media metrics such as the number of "likes" or "retweets" received by the tweet.

For our dependent measure, we identified the 40 claims that were affirmed or negated in each fact check (see Figure 1). Claims were presented to participants without any social-media-based formatting during the truth rating phase (e.g., "A Florida school ditched Common Core and then soared to No. 1."). Full materials are available at the project's OSF site: https://osf.io/ndru4/?view_only=ca9c1209c94f449ea5ca029babf1fadb

For Experiment 1, we used 32 of the 40 total fact checks we had selected. This ensured that the number of items used was evenly divisible by the number of cells in the design (16: two

levels for each the four independent variables—claim veracity, fact check exposure, fact check format, and fact check origin). We created two sets of fact checks to balance the "journalist" and "researcher" generated versions of each fact check. Each set contained four journalist and researcher generated fact checks in each of the four formats (true before, true after, false before, false after), for a total of 32 unique fact checks (16 journalist created and 16 researcher created). Each set was then subdivided into two subsets of 16 fact checks total (2 from each format), to counterbalance which claims were new or fact-checked. Participants were then randomly assigned to one of the two sets of 32 fact checks and to view one of the two subsets of 16 fact checks during the exposure phase. As mentioned above, we arbitrarily assigned levels of fact check format and origin to "new claims" for comparison purposes (2 per format).

Procedure.

Fact check exposure phase. After giving informed consent, participants entered their age and were then told that we were interested in their opinions on tweets that fact-checked true and false claims. Participants were not instructed that they would be asked to remember the content of the tweets. They then saw a series of 16 fact-checking tweet screenshots in a random order and were asked to indicate how interesting each tweet appeared on a 6-point scale (1 = Very Interesting, 2 = Interesting, 3 = Slightly Interesting, 4 = Slightly Uninteresting, 5 = Uninteresting, 6 = Very Uninteresting).

Claim rating phase. Immediately after the fact check exposure phase, participants were instructed to rate the truth of statements, some of which were related to the tweets they previously saw. The instructions also indicated that some statements would be true and some would be false. Participants then saw the full series of 32 claims in a random order and were asked to indicate how true or false each claim appeared on a 6-point scale (1 = Definitely False,

2 = Probably False, 3 = Possibly False, 4 = Possibly True, 5 = Probably True, 6 = Definitely True). Finally, participants were informed about the purpose of the study, and asked not to share information about the study with other potential participants.

Results.

Pre-registered analysis plans and hypotheses are available at the project's OSF site: https://osf.io/ndru4/?view_only=ca9c1209c94f449ea5ca029babf1fadb. All model-fitting and statistical analyses were conducted using the "lme4" package (Bates, Mächler, Bolker, & Walker, 2015) in the R programming language (R Core Team, 2020). All statistical tests were conducted at the .05 alpha level.

Effects of exposure to fact-checking tweets on rating of claims. We first examined the impact of exposure to fact checks on participants' truth ratings of true and false claims. Our question was whether participants would more accurately rate claims as "true" or "false" after having seen a fact check as compared to when they were viewing the claim for the first time. As shown in Figure 2, the difference in truth ratings for true and false claims was greater for fact-checked claims than for new claims. Our second question was whether these effects would vary based on the format of that fact check. As shown in Figure 2, they did not —truth ratings for fact-checked true and false claims were very similar across the before and after formats.

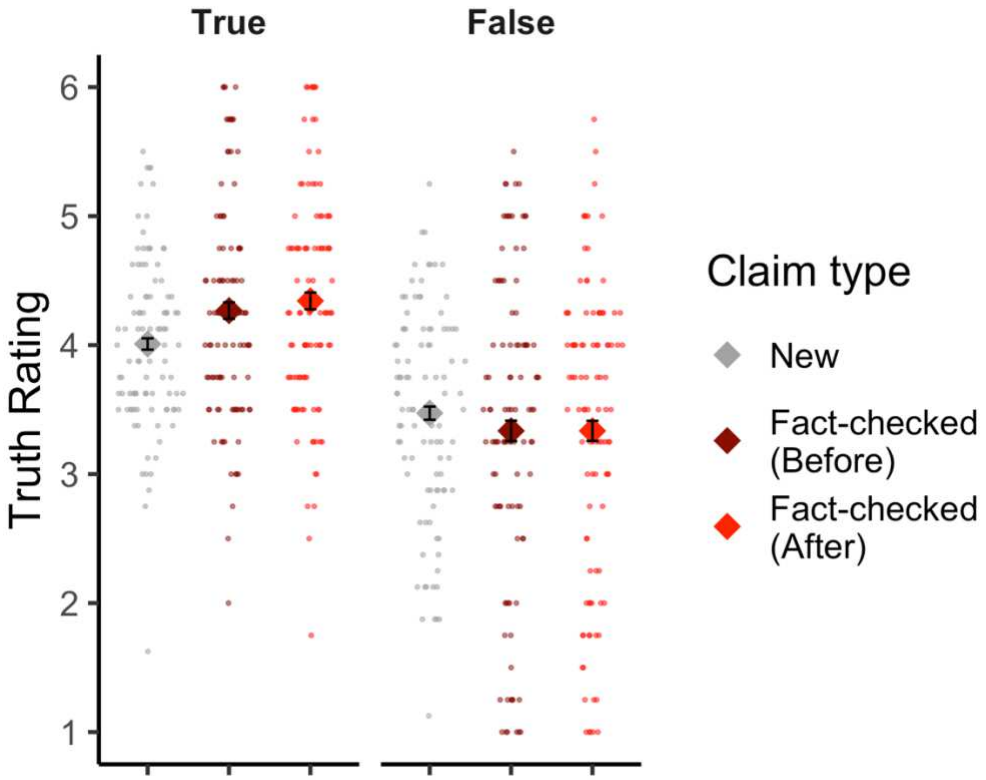

Figure 2. Mean truth ratings for true and false claims that were new as well as for those that were initially encountered during the fact check exposure phase, split by the fact check's format. Rating scale ranged from 1 = Definitely False to 6 = Definitely True. Each dot represents one participant (N = 100), with values horizontally shifted to represent the density distribution. Diamonds reflect group means and error bars reflect the standard error of the mean.

Following our preregistration, we evaluated these data statistically by fitting a mixed-effects linear regression model to the truth rating data. The model included claim veracity (true = 0.5, false = -0.5), fact check exposure (fact-checked claim = -0.5, new claim = 0.5), fact check format (evaluation before claim = -0.5, evaluation after claim = 0.5), and all possible interactions as fixed effects. As preregistered, we used the "buildmer" package in R (Voeten, 2019) to identify the model with the maximal random effects structure that avoided convergence errors.

For this step, possible random effects terms were added in ascending order of the significance of a likelihood-ratio test of the term until the model no longer converged when fit using the "bobyqa" optimizer. Possible by-claim random effects terms for our design included intercepts and effects of exposure, format, and their interaction (but not veracity, as each claim was either true or false). Possible by-subject random effects terms for our design included intercepts and effects of claim veracity, fact check exposure, fact check format and all possible interactions. Note that we randomly pre-assigned a "fact check format" level for each "new" claim for which participants were not actually exposed to a fact check, as described above. Thus, the "fact check format" variable becomes crossed with (as opposed to nested within one level of) fact check exposure. As a result, the model includes a main effect of format, as well as exposure-format interaction terms. The final random effects structure after this procedure, as well estimates produced after fitting this model to the 3,200 truth ratings, are provided in Table 1.

Table 1.

Mixed-effects model for Experiment 1 with fixed effects of claim veracity, fact check exposure, fact check format, and all interactions, as well as random slopes and intercepts.

Fixed Effects	Estimate	*SE*	*Df*	*t* Value	*p* Value
Intercept	3.781	0.065	100.2	57.907	<.001
Claim veracity	**0.753**	**0.122**	**92.24**	**6.178**	**<.001**
Fact check exposure	-0.079	0.044	97.76	-1.799	.075
Fact check format	-0.003	0.055	32.27	-0.067	.947
Veracity*Exposure	**-0.428**	**0.084**	**2,744**	**-5.103**	**<.001**
Exposure*Format	-0.092	0.084	2,741	-1.097	.273
Veracity*Format	-0.042	0.103	2.870	-0.408	.686
Veracity*Exposure*Format	-0.226	0.168	2,741	-1.343	.179

Random Effects	Variance	*SD*	Correlation		
Participant (Intercept)	0.271	0.520			
Participant (Veracity)	0.864	0.929	0.42		
Participant (Format)	0.044	0.209	0.22	0.08	
Participant (Exposure)	0.017	0.130	0.06	-0.14	-0.09
Claim (Intercept)	0.036	0.189			
Claim (Format)	0.028	0.167	0.23		

Note. Model was fit to 3,200 truth ratings from 100 participants across 32 claims. Bolded values indicate significant effects. Correlation values in each row reflect correlations between the term and all preceding random effects terms in the same level (participant or claim) in order of appearance in the table.

As shown in Table 1, we found a significant fixed effect of claim veracity, such that true claims ($M = 4.16$) were estimated to be rated 0.753 points higher than false claims ($M = 3.40$) on the 6-point rating scale (1 = Definitely False, 6 = Definitely True). This effect was qualified by a significant interaction between claim veracity and fact check exposure. Follow-up *t*-tests revealed that, participants rated true claims ($M = 4.01$) as more "true" than false claims ($M = 3.47$) both when they were new ($t(1598) = 7.95$, $p < .001$, 95% CI of the difference [0.40, 0.67], $d = 0.40$), and when they had previously seen the fact-checking tweet (true $M = 4.31$, false $M = 3.34$; $t(1598) = 13.49$, $p < .001$, 95% CI of the difference [0.83, 1.11], $d = 0.67$. However, the difference was larger for fact-checked claims (non-overlapping 95% CI for the estimate of the difference). The interaction effect between veracity, exposure, and format was non-significant; the effects of exposure on ratings for true and false claims did not differ based on how the fact check was worded. No other fixed effects were significant. Overall, participants were able to distinguish between true and false claims even without having seen any fact checks. However, exposure to fact checks increased participants' accuracy at rating claims as "true" or "false".

Effects of fact check origin. Recall that, for each fact check, we created a second, researcher-generated version to ensure that fact check format was not confounded with the content or topic of the fact checks. As a pre-registered exploratory analysis, we examined whether the researcher-originated and journalist-originated versions of the fact checks influenced participants' truth ratings in similar or different ways. This analysis was conducted using a mixed-effects linear regression model constructed in the exact same manner as the main analysis reported above, with the addition of fact check origin (journalist, researcher) and all possible interactions as fixed effects, as well as random effects of fact check origin and respective interactions by-claim and by-participant. Neither the main effect of fact check origin, nor any of the additional interaction effects involving fact check origin were significant (all ps > .05). Full results are listed as supplementary analyses on the project's OSF site: https://osf.io/ndru4/?view_only=ca9c1209c94f449ea5ca029babf1fadb

Experiment 2

Experiment 1 established that exposure to fact checks improved the accuracy of people's beliefs regarding the verified claims in an immediate rating task, and that this effect was similar for both types of fact check formats (i.e., evaluations before or after the claims). The goal of Experiment 2 was to examine how people retain the information in these fact checks over a week-long delay, and whether this retention varies based on the format of the original fact check. Participants read a series of fact checks and rated the claims either immediately and after a week-long delay, or only after a week-long delay. Critically, we predicted that participants would begin to forget the falsehood of negated claims more often when the evaluation was placed after the claim than when it was placed before, as the latter type of negation would produce stronger associations between the claim and its status as false in memory.

Method.

Participants. One hundred and two adult participants (M_{age} = 39.20, SD = 12.46) were recruited from Amazon's Mechanical Turk Platform (MTurk) to complete the first session of this study online through the Finding Five platform (FindingFive Team, 2019) for a payment of $3.63. We restricted our sample to participants from the United States who had at least a 95% approval rating on MTurk and who had not completed similar studies from our lab in the past. The number of participants exceeded our pre-registered sample size by two because two participants were "in progress" when our requested number of participants was reached and finished the study afterwards.

Ninety-one of the original 102 participants (M_{age} = 38.27, SD = 12.20) returned to complete the second session of this study for a payment of $1.81. Following our pre-registration, we excluded the 11 participants who did not return, only analyzing data for these 91 participants. 38 of the included participants were in the repeated rating condition, and 53 were in the delayed-only rating condition.

Design. The experiment had a 2 (rating condition: repeated (immediate & one-week delayed), single (one-week-delayed only)) × 2 (claim veracity: true, false) × 2 (fact check format: evaluation before claim, evaluation after claim) × 2 (fact check origin: journalist, researcher) mixed design, with rating condition being a between-subjects factor, and all other factors manipulated within-subjects, as shown in Figure 3. The critical comparison was between immediate ratings from participants in the repeated rating condition with delayed ratings from participants in the single rating condition. This allows us to compare participants' initial ratings at different delays while also ensuring participant groups are similar (i.e., both groups took part

in a two-session study).

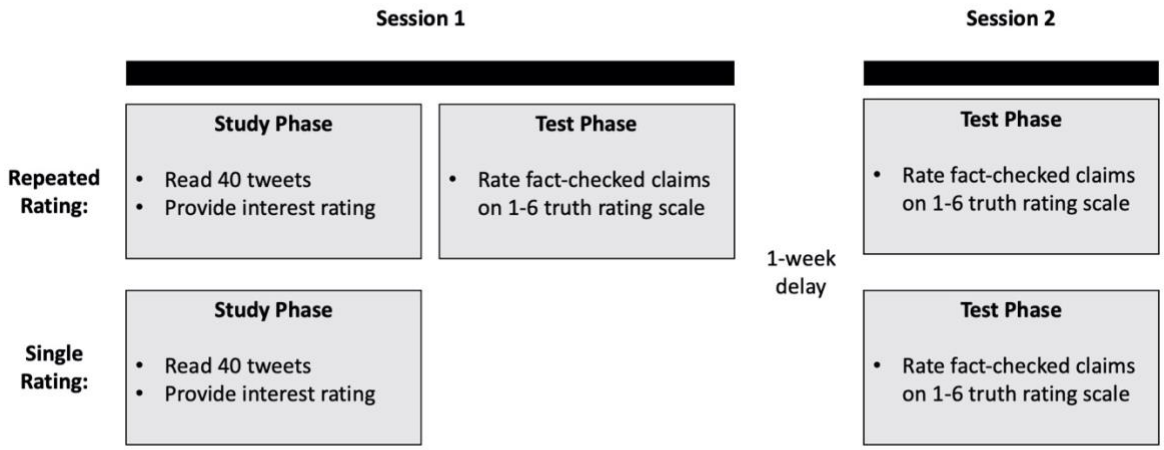

Figure 3. Study design for Experiment 2.

Materials. The full set of 40 claims and their corresponding fact-checking tweets described in the Materials section of Experiment 1 were used in this experiment. As in Experiment 1, we assigned participants to one of two fact check stimulus sets for counterbalancing purposes. Each set contained either the "journalist" or the "researcher" generated version of each of the 40 fact-checks. The fact-checks were distributed evenly such that each set contained five journalist and researcher generated fact checks for true and false claims in the before and after format. Note that no claims were "new" in this experiment—participants saw all 40 fact checks in their stimulus set.

Procedure.

Fact check exposure phase. After giving informed consent, participants entered their age and were then told that we were interested in their opinions on tweets that fact-checked true and

false claims. Participants were not instructed that they would be asked to remember the content of the tweets. Participants then saw a series of 40 fact-checking tweet screenshots in a random order and were asked to indicate for each tweet how interesting the tweet appeared on a 6-point scale (1 = Very Interesting, 2 = Interesting, 3 = Slightly Interesting, 4 = Slightly Uninteresting, 5 = Uninteresting, 6 = Very Uninteresting). Then, participants in the single rating condition were thanked and asked to await a study invitation in one week. Participants in the repeated rating condition proceeded to the immediate claim rating phase.

Immediate claim rating phase. Immediately after the fact check exposure phase, participants in the repeated rating condition were instructed that they were to rate the truth of statements related to the tweets they previously saw. The instructions also indicated that some statements would be true and some would be false. Participants then saw a series of 40 claims in a random order and were asked to indicate for each claim how true or false the claim appeared on a 6-point scale (1 = Definitely False, 2 = Probably False, 3 = Possibly False, 4 = Possibly True, 5 = Probably True, 6 = Definitely True). Finally, participants were thanked and asked to await a study invitation in one week.

Delayed claim rating phase. One week after the previous phases, the second session of the study became open to all initial participants and closed 2 days later. This session consisted only of the delayed claim rating phase. All participants were instructed that they were to rate the truth of statements related to the tweets they saw the previous week. Participants then saw a series of 40 claims in a random order and were asked to indicate for each claim how true or false the claim appeared on a 6-point scale (1 = Definitely False, 2 = Probably False, 3 = Possibly False, 4 = Possibly True, 5 = Probably True, 6 = Definitely True). Finally, participants were

informed about the purpose of the study, and asked not to share information about the study with other potential participants.

Results.

Effects of fact-checking tweets over time. The key question we examined was whether the format of the fact-checking tweet would affect participants' abilities to accurately rate claims as "true" or "false" over time. As shown in Figure 4, it did not. Participants consistently rated true claims as more "true" than false claims whether they were tested immediately or one week after exposure to fact checks of the claim. In addition, this pattern of ratings did not vary based on how the original fact check was formatted.

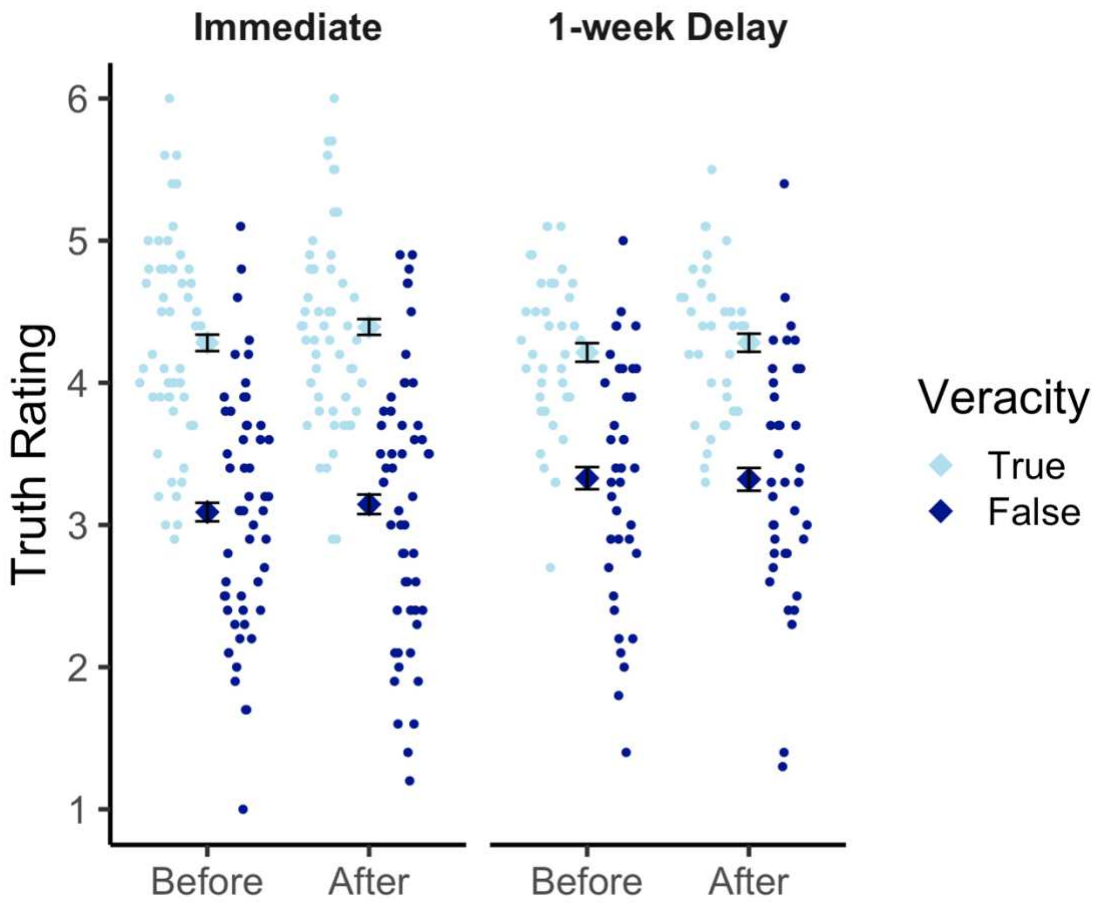

Figure 4. Mean truth ratings for claims rated immediately and after a one-week delay, split by fact check format. Rating scale ranged from 1 = Definitely False to 6 = Definitely True. Each dot represents one participant ($N = 91$), with values horizontally shifted to represent the density distribution. Diamonds reflect group means and error bars reflect the standard error of the mean.

Following our preregistration, we evaluated these data statistically using a mixed-effects linear regression model for immediate truth ratings by participants in the repeated rating condition, and delayed truth ratings by participants in the single rating condition. The model had claim veracity (true = 0.5, false = -0.5), fact check format (evaluation before claim = -0.5, evaluation after claim = 0.5), time (immediate test = -0.5, one-week delayed test = 0.5), and all possible interactions as fixed effects. The pre-registered approach to the random effects structure only specified that we would adopt the maximal model that still converged. For the sake of consistency with other experiments reported in this manuscript, we describe the results obtained using the "buildmer" package as described in Experiment 1. Possible by-claim random effects terms for our design included intercepts and effects of time, fact check format, and their interaction (but not veracity, as each claim was either true or false). Possible by-subject random effects terms for our design included intercepts and effects of claim veracity, fact check format and their interactions. The final random effects structure as well as the fixed-effect estimates produced after this model was fit to the 3,640 truth ratings are provided in Table 2 below.

Table 2.

Mixed-effects model for Experiment 2 with fixed effects of claim veracity, fact check format, time, and all interactions as well as random slopes and intercepts.

Fixed Effects	Estimate	*SE*	*Df*	*t* Value	*p* Value
Intercept	3.757	0.058	99.10	64.89	<.001
Claim veracity	**1.071**	**0.130**	**103.9**	**8.229**	**<.001**

Fact check format	0.055	0.042	3,385	1.330	.184
Time	0.058	0.102	85.71	0.576	.566
Veracity*Time	-0.297	0.236	87.92	-1.256	.212
Veracity*Format	0.067	0.083	3,385	0.806	.420
Format*Time	-0.054	0.083	3,388	-0.643	.520
Veracity*Format*Time	0.021	0.167	3,388	0.124	.901
Random Effects	Variance	*SD*	Correlation		
Participant (Intercept)	0.187	0.432			
Participant (Veracity)	1.058	1.028	0.37		
Claim (Intercept)	0.032	0.180			
Claim (Time)	0.010	0.100	-0.67		

Note. Model was fit to 3,640 truth ratings from 91 participants across 40 claims. Bolded values indicate significant effects. Correlation values in each row reflect correlations between the term and all preceding random effects terms in the same level (participant or claim) in order of appearance in the table.

As shown in Table 2, we found a significant fixed effect of claim veracity, such that true claims (M = 4.30) were estimated to be rated 1.07 points higher than false claims (M = 3.20) on the 6-point rating scale (1 = Definitely False, 6 = Definitely True). Further, contrary to our predictions, we found neither a significant main effect of fact check format, nor any significant interaction effects involving format; participants' overall pattern of ratings was unaffected by the format of the fact check for each claim. However, we also did not find a main effect of time, or an interaction between time and veracity, suggesting that participants may not have been reliably forgetting the truth of any of the claim and preventing us from examining differences in forgetting based on how the fact check was originally worded.

Effects of fact check origin. As in Experiment 1, we also explored whether the trends observed in our main analysis differed for journalist-originated and researcher-originated versions of each fact check by adding fact check origin as a predictor to our main mixed-effects linear regression model. We did not observe a significant main effect of fact check origin, or any interaction effects involving fact check origin except for an unhypothesized three-way interaction between claim veracity, time, and fact check origin. Journalist-generated fact checks were slightly more effective than researcher-generated tweets on the immediate test (higher ratings for true claims and lower ratings for false claims), but efficacy was similar on the delayed test. Due to difficulties in interpretation of higher-order interactions and due to the absence of significant lower-level two-way interactions pertaining to this effect, we refrain from commenting further on this specific effect. However, full results of this analysis are available on the project's OSF site: https://osf.io/ndru4/?view_only=ca9c1209c94f449ea5ca029babf1fadb

Experiment 3

Experiment 2 demonstrated that fact checks containing evaluations before and after the claim are remembered similarly over time. However, after the week-long delay, there was little evidence that participants were forgetting which claims were true and which were false, suggesting that differences in forgetting may only be apparent over longer time scales. Thus, Experiment 3 is a direction replication of Experiment 2 with a longer delay of 3 weeks, designed to determine whether fact checks of different formats, when forgotten, are forgotten at different rates over time. Again, we hypothesized that participants would begin to forget the falsehood of negated claims more often when the evaluation was placed after the claim than when it was placed before.

Method.

Participants. Two hundred and fifty-two adult participants (M_{age} = 37.67, *SD* = 10.18, one participant not reporting) were initially recruited in the same manner as described in Experiment 2. This number of participants recruited matched our pre-registered sample size for initial recruitment, which was determined by a simulation-based power analysis conducted using the simR package (Green & MacLeod, 2016). This power analysis was designed to determine the sample size needed to detect a minimal delay-format interaction of interest for the false claims, corresponding to our hypothesis that evaluative tags are forgotten differently over time based on the original format of the fact check.

In the absence of heuristics for small, medium, and large standardized coefficients, our selection of the minimal effect size of interest was based on the observed effects in Experiment 1. In that study, we observed a partially standardized veracity-exposure interaction effect of 0.30. That is, one estimate of the effect size of the efficacy of these fact-checking tweets is 0.30. We reasoned that fact check format would have a meaningful effect on truth ratings if the effect was approximately half of this value, so we selected 0.15 as the minimal effect size of interest for the power analysis.

The power analysis used simR to simulate data based on the actual data from the false claims in Experiment 2, with the modification that the delay-format interaction term was set at 0.15. The results of 1,000 simulation runs indicated that a sample size of 198 participants in total would be needed to provide 80% power to detect this effect. Anticipating a 20% rate of attrition, as observed in Experiment 2, and rounding up to account for the number of counterbalancing conditions, we preregistered an initial sample size of 252.

One hundred and seventy-eight of the original 252 participants (M_{age} = 38.26, *SD* = 10.55, one participant not reporting) returned to complete the second session of this study for a

payment of $1.81. Following our pre-registration, we excluded the 74 participants who did not return, only analyzing data for these 178 participants, which put our sample size slightly short of our 198-person goal. 89 of these participants were in the repeated rating condition, and 89 were in the delayed-only rating condition.

Procedure. The design, materials, and procedure used for Experiment 3 are identical to those of Experiment 2, with one exception: the delayed rating session became open to participants 3 weeks (as opposed to 1 week) after the initial session and remained open for 6 days afterwards.

Results.

Effects of fact-checking tweets over time. As in Experiment 2, the key question was whether the format of the fact-checking tweet would affect participants' abilities to accurately rate the claims as "true" or "false" over time. As shown in Figure 5, we do not observe any effects of format. Participants consistently rated true claims as more "true" than false claims at both time points, and this difference was larger for participants tested immediately, suggesting that participants were forgetting the truth of each claim over time. However, this pattern of ratings was very similar across the before and after formats.

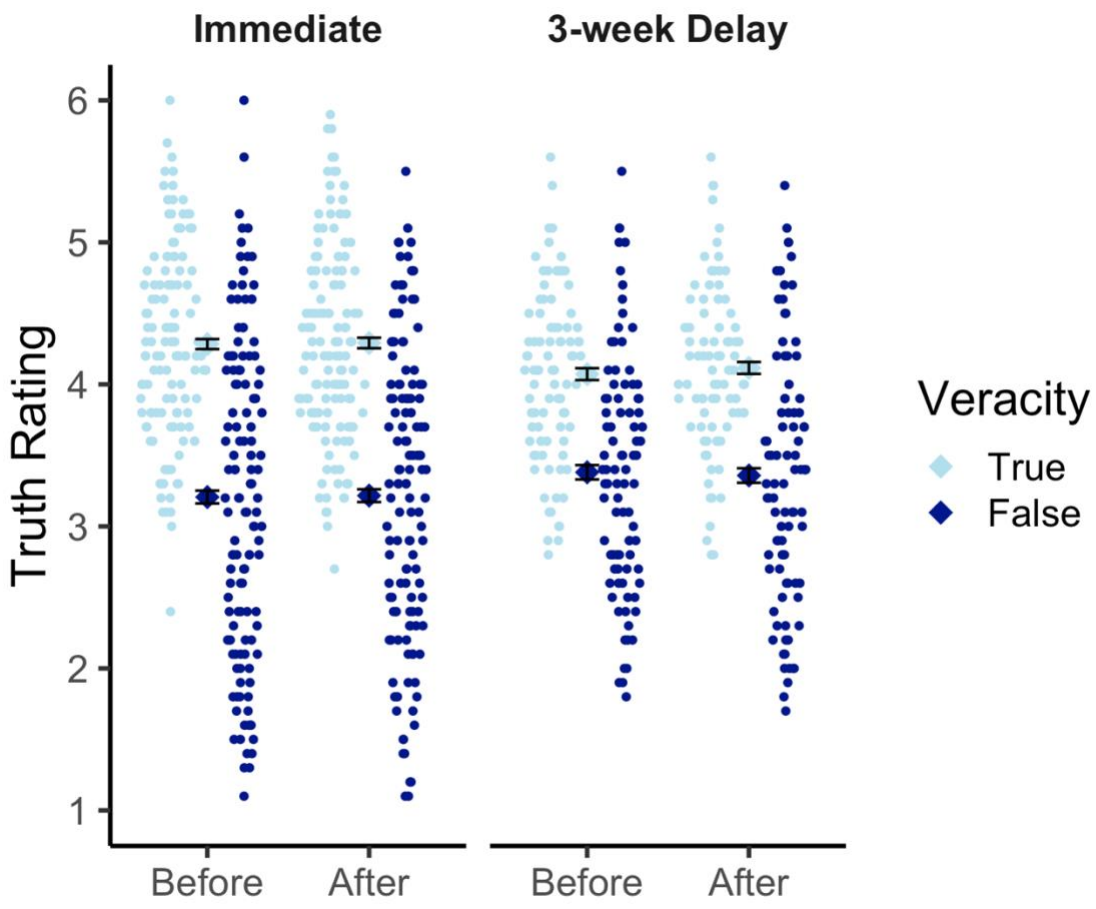

Figure 5. Mean truth ratings for claims rated immediately and after a three-week delay, split by fact check format. Rating scale ranged from 1 = Definitely False to 6 = Definitely True. Each dot represents one participant (N = 178), with values horizontally shifted to represent the density distribution. Diamonds reflect group means and error bars reflect the standard error of the mean.

Following our preregistration, we evaluated these data statistically using a mixed-effects linear regression model for immediate truth ratings by participants in the repeated rating condition and delayed truth ratings by participants in the single rating condition. The model had claim veracity (true = 0.5, false = -0.5), fact check format (evaluation before claim = -0.5,

evaluation after claim = 0.5), time (immediate test = -0.5, three-week delayed test = 0.5), and all possible interactions as fixed effects. As pre-registered, we used the "buildmer" package as described in Experiment 2 (including the same maximal possible random effects structure) to identify the maximal feasible model. The final random effects structure, as well fixed-effect estimates produced after this model to the 7,120 truth ratings provided in Table 3 below.

Table 3.

Mixed-effects model for Experiment 3 with fixed effects of claim veracity, fact check format, time, and all interactions as well as random slopes and intercepts.

Fixed Effects	Estimate	*SE*	*Df*	*t* Value	*p* Value
Intercept	3.702	0.057	103.1	64.71	<.001
Claim veracity	**0.949**	**0.117**	**110.1**	**8.094**	**<.001**
Fact check format	0.010	0.029	175.1	0.350	.727
Time	0.060	0.088	171.4	0.680	.498
Veracity*Time	**-0.449**	**0.183**	**175.8**	**-2.449**	**.015**
Veracity*Format	0.028	0.057	6,527	0.498	.619
Format*Time	0.002	0.059	175.0	0.032	.974
Veracity*Format*Time	0.070	0.114	6,524	0.611	.542
Random Effects	Variance	*SD*	Correlation		
Participant (Intercept)	0.256	0.506			
Participant (Veracity)	1.138	1.067	-0.46		
Participant (Format)	0.009	0.094	-0.07	-0.35	
Claim (Intercept)	0.065	0.255			
Claim (Time)	0.047	0.218	0.63		

Note. Model was fit to 7,120 truth ratings from 178 participants across 40 claims. Bolded values indicate significant effects. Correlation values in each row reflect correlations between the term and all preceding random effects terms in the same level (participant or claim) in order of appearance in the table.

As shown in Table 3, we found a significant effect of claim veracity, such that true claims (M = 4.18) were estimated to be rated 0.949 points higher than false statements (M = 3.23) on the 6-point rating scale (1 = Definitely False, 6 = Definitely True). This effect was qualified by a significant interaction between claim veracity and time. Follow-up t-tests on the truth rating data revealed that, participants rated true claims (M = 4.26) as more "true" than false claims (M = 3.09) on the immediate test ($t(3558) = 23.56$, $p < .001$, 95% CI of the difference [1.08, 1.27], d = 0.79). A similar trend was observed on the three-week delayed test, though the 95% confidence interval for the estimate of the difference was smaller than and nonoverlapping with that of the immediate test (true M = 4.09, false M = 3.37; $t(3358) = 15.613$, $p < .001$, 95% CI of the difference [0.63, 0.82], d = 0.52). Thus, we see evidence that participants are beginning to forget which claims are true and which are false over this longer delay. However, we did not observe a main effect of fact check format, nor any interaction effects involving fact check format, suggesting that the above pattern of ratings are consistent regardless of the phrasing of the initial tweet.

Effects of fact check origin. As in prior experiments, we also explored whether the trends observed in our main analysis differed for journalist-originated and researcher-originated versions of each fact check by adding fact check origin as a predictor to our main mixed-effects linear regression model. We observed a small main effect of fact-check origin such that participants rated claims as 0.06 points more true after reading a fact check from the researcher compared to when they read an original fact check ($M_{original}$ = 3.67, $M_{researcher}$ = 3.73; $t(6525)$ = 2.17, p =.030). Note, however, that this main effect was calculated across both true and false claims. Thus, this finding does not imply that the original fact checks were less effective. We also observed a three-way interaction between claim veracity, fact check format, and fact check

origin. False claims were given higher ratings for the "after" than the "before" journalist-generated fact checks, and the opposite pattern held for researcher-generated fact checks. However, true claims were rated relatively similarly regardless of fact check origin and format. Thus, the false claims that journalists choose to write in the "after" format may be naturally more believable than claims chosen for the "before" format. However, it is important to note that this interaction (as well as the main effect of fact check origin) were not significant in either of the previous studies. No other interaction terms involving fact check origin were significant. Full results of this analysis are available on the project's OSF site: https://osf.io/ndru4/?view_only=ca9c1209c94f449ea5ca029babf1fadb.

General Discussion

The present research examined the effects of exposure to two different kinds of short, social-media-based fact checks on people's beliefs. Consistent with past literature (Ecker, Lewandowsky, et al., 2020; Ecker, O'Reilly, et al., 2020), these fact checks were effective. In Experiment 1, participants were better able to distinguish true from false claims after exposure to relevant fact checks, and in Experiment 2, this discrimination ability remained consistent even after a week-long delay. However, contrary to our predictions, there was no difference in the efficacy of fact checks which placed an evaluation before the claim (e.g., "X did not do Y, as A claims") or after the claim (e.g., "A claims X did Y. No, this is false") is stated in its entirety. In Experiment 3, participants were less able to discriminate true and false claims after a three-week delay, suggesting they began to forget the information conveyed in the fact checks. However, this forgetting was similar across both fact check formats. In other words, we did not observe any evidence that one format of fact check was more effective.

From a cognitive psychological perspective, the present work offers two main contributions. First, Experiment 1 adds to the growing literature demonstrating that repeating claims in order to refute them is unlikely to inadvertently increase belief in the claim (see Swire-Thompson et al., 2020 for a review). Repetition of true and false claims by themselves can heighten belief (Dechêne, Stahl, Hansen, & Wänke, 2010), but indicating that the claim is false on first exposure can prevent these harms. Critically, unlike in prior studies (e.g., Ecker, Lewandowsky, et al., 2020), this pattern was demonstrated while obtaining truth judgements in a mixed list of claims that were repeated in a prior fact check and new claims for which no fact check was seen. These mixed lists often exaggerate the effects of repetition on belief by providing a reference point in the unrepeated claims (Dechêne et al., 2009). Thus, Experiment 1 provides stronger evidence than has been previously shown against the idea that short fact checks may adversely increase belief in falsehoods.

Second, the present work demonstrates that, contrary to our hypothesis, modifying the location of the negative operator in a negated message has no impact on memory for the content of that message. While participants began to forget which claims were affirmed and which were negated after a delay of three weeks in Experiment 3, this forgetting was similar regardless of the format of the original fact check. While this finding has implications for theories of negation processing, the exact reason for the lack of observed differences is a question for future research.

One possibility is that the predicted dynamics do not reliably play out on the short, one or two-sentence timescale of negated messages. For instance, any predicted increase in distrustful encoding processes may not reliably be elicited by a “no” at the beginning of a sentence negating a false claim, relative to the claim followed by statement that it is false. Similarly, the increased proximity of the negative operator to the main verb of the claim in the “before” format may only

have negligible effects on the durability of the encoded association. The same logic may explain why placing the negation after the claim did not improve memory either. Placing the evaluation in the sentence immediately after the claim may not provide enough time for participants to generate a prediction about the truth of the statement, and so no feedback-based benefits may have been observed. Instead, the negated messages examined here may be short enough that they give rise to mental associations between a claim and its truth status that are of similar strength regardless of the specific locations of the corresponding words or phrases relative to one another.

However, an alternate explanation is that multiple of these cognitive mechanisms may be involved, resulting in similar memory performance for both formats, but for different reasons. For example, placing the negation before the claim may improve memory by triggering heightened elaboration regarding the truth status of the claim upon initial reading, and placing the negation after the claim may improve memory by emphasizing differences between the predicted and actual truth status of a claim. If these effects are comparable in magnitude, we would expect to see similar performance across formats, despite different processing of the initial stimuli. Future work is needed to address the extent to which negated messages of different formats are processed similarly.

Regardless of the exact mechanisms involved in processing negated messages of different formats, our findings have clear practical implications. Overall, this work demonstrates that current strategies being employed to communicate fact checks on social media are effective. Exposure to the fact check stimuli used in the current work, which were drawn from PolitiFact and subject to only minor wording modifications, improved people's ability to discern truths from falsehoods, an effect that remained stable for at least one week. In addition, the equivalent effects for both fact check formats indicate that multiple paths are being effectively employed to

communicate the truth of questionable claims on social media. We suggest that fact-checkers should continue to use short-format fact checks in their work without worrying about the specific phrasing.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2009). Mix me a list: Context moderates the truth effect and the mere-exposure effect. *Journal of Experimental Social Psychology*, *45*(5), 1117–1122. https://doi.org/10.1016/j.jesp.2009.06.019

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

Ecker, U. K. H., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 41. https://doi.org/10.1186/s41235-020-00241-6

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

Ecker, U. K. H., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*, *111*(1), 36–54. https://doi.org/10.1111/bjop.12383

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88–92. https://doi.org/10.3758/PBR.16.1.88

FindingFive Team. (2019). *FindingFive: A web platform for creating, running, and managing your studies in one place.* NJ, USA: FindingFive Corporation (nonprofit). Retrieved from https://www.findingfive.com

Fridkin, K., Kenney, P. J., & Wintersieck, A. (2015). Liar, liar, pants on fire: How fact-checking influences citizens' reactions to negative advertising. *Political Communication*, *32*(1), 127–151. https://doi.org/10.1080/10584609.2014.914613

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

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

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Mayo, R. (2019). Knowledge and distrust may go a long way in the battle with disinformation: Mental processes of spontaneous disbelief. *Current Directions in Psychological Science*, *28*(4), 409–414. https://doi.org/10.1177/0963721419847998

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology*, *40*(4), 433–449. https://doi.org/10.1016/j.jesp.2003.07.008

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*, 939–960. https://doi.org/10.1007/s11109-019-09528-x

PolitiFact. (2020, February 22). No, the UN is not planning to implant the world with biometric IDs. Read more here: Http://bit.ly/38JZDlk [Tweet]. Retrieved May 18, 2020, from https://twitter.com/PolitiFact/status/1231252307095932929

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Schul, Y. (1993). When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology*, *29*(1), 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*(5), 668–679. https://doi.org/10.1037/0022-3514.86.5.668

Stencel, M., & Luther, J. (2021). *Fact-checking census shows slower growth*. Retrieved from https://reporterslab.org/fact-checking-census-shows-slower-growth/

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*. Advance online publication. https://doi.org/10.1016/j.jarmac.2020.06.006

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, *28*(3), 247–253. https://doi.org/10.1177/0963721419827854

Voeten, C. C. (2019). *buildmer: Stepwise elimination and term reordering for mixed-effects regression*.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation: emotions and misperceptions. *Journal of Communication*, *65*(4), 699–719. https://doi.org/10.1111/jcom.12164