British Journal of Psychology (2019)

The British Psychological Society

www.wileyonlinelibrary.com

The effectiveness of short-format refutational fact-checks

Ullrich K. H. Ecker*, Ziggy O'Reilly, Jesse S. Reid and Ee Pin Chang

School of Psychological Science, University of Western Australia, Perth, Western Australia, Australia

Fact-checking has become an important feature of the modern media landscape. However, it is unclear what the most effective format of fact-checks is. Some have argued that simple retractions that repeat a false claim and tag it as false may backfire because they boost the claim's familiarity. More detailed refutations may provide a more promising approach, but may not be feasible under the severe space constraints associated with social-media communication. In two experiments, we tested whether (1) simple 'false-tag' retractions can indeed be ineffective or harmful; and (2) short-format (140-character) refutations are more effective than simple retractions. Regarding (1), simple retractions reduced belief in false claims, and we found no evidence for a familiarity-driven backfire effect. Regarding (2), short-format refutations were found to be more effective than simple retractions after a 1-week delay but not a one-day delay. At both delays, however, they were associated with reduced misinformation-congruent reasoning.

Fact-checking is now recognized as an important tool in the contemporary media landscape of a democratic society (Graves, 2017). Ideally, fact-checking can help guide public discourse and foster evidence-based decision making and policy creation, as well as shape perceptions of political candidates (Fridkin, Kenney, & Wintersieck, 2015; Gottfried, Hardy, Winneg, & Jamieson, 2013; Wintersieck, 2017; Wintersieck, Fridkin, & Kenney, 2018; also see Bode & Vraga, 2015; Cobb, Nyhan, & Reifler, 2013). While some have questioned the impact of fact-checking given the scope of the misinformation problem (Lazer *et al.*, 2018; Lewandowsky, Ecker, & Cook, 2017; Shao *et al.*, 2018; Swire, Berinsky, Lewandowsky, & Ecker, 2017), the impact is almost certainly a positive one (Nyhan & Reifler, 2015a; obviously, the net societal benefit of fact-checking is difficult to determine, as we do not know what the world would currently look like without fact-checking). Even the mere threat of potentially being fact-checked can reduce the deliberate dissemination of disinformation (as demonstrated in U.S. state legislators; Nyhan & Reifler, 2015b).

However, the fact-checking movement does indeed face issues. One issue relates to the perceived objectivity of fact-checking sources (Shin & Thorson, 2017; Stencel, 2015). Clearly, fact-checking can only have broad impact if a large majority view the fact-

**Correspondence should be addressed to Ullrich Ecker, University of Western Australia (M304), 35 Stirling Hwy, Perth 6009, WA, Australia (email: ullrich.ecker@uwa.edu.au).*

The data relating to this study are available at https://osf.io/c45tf/?view_only=ca57d7349b07453cabe94feb6d72b638

DOI:10.1111/bjop.12383

checking source as objective and neutral, and the fact-checks themselves as unbiased, fair, and grounded in reliable evidence (Brandtzaeg & Følstad, 2017; Guillory & Geraci, 2013). On social media, fact-checks are also more likely to be accepted from known sources rather than strangers (Margolin, Hannak, & Weber, 2018).

A more fundamental issue is that – even if the fact-checker is perceived as objective and the consumers themselves are unbiased – fact-checking communications appear to be less effective than desired (Berinsky, 2015; Garrett, Nisbet, & Lynch, 2013; Nyhan & Reifler, 2010; Thorson, 2016). This is corroborated by experimental psychological work that has consistently shown that corrected misinformation continues to influence people's memory and reasoning even if the correction is understood and remembered – a phenomenon termed the continued influence effect of misinformation (Johnson & Seifert, 1994; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Seifert, 2002). Moreover, it is unclear what fact-checking format is most effective (Amazeen, Thorson, Muddiman, & Graves, 2018; Young, Jamieson, Poulsen, & Goldring, 2018). The most common format used by fact-checking websites or social-media accounts devoted to fact-checking restates the misinformation while adding some variant of a 'false' tag (see Figure 1). Most fact-checks will point to additional information justifying the labelling, but especially on social media, consumers will often be exposed only to the false claim and the corresponding tag, and will have to actively seek out additional information, for example by following a hyperlink.

From a theoretical point of view, some have argued that repeating misinformation in this manner should be avoided because such corrections increase the misinformation's familiarity, which might have undesired consequences: The more familiar information is, the easier it is retrieved from memory, and the more likely it is accepted as true (Dechêne, Stahl, Hansen, & Wanke, 2010; Weaver, Garcia, Schwarz, & Miller, 2007); thus, boosting misinformation familiarity might counteract and offset the intended effect of the correction, potentially even leading to ironic backfire effects (Lewandowsky *et al.*, 2012; Peter & Koch, 2016; Skurnik, Yoon, Park, & Schwarz, 2005; also see Swire, Ecker, & Lewandowsky, 2017). Moreover, if communication recipients have not encountered a particular false claim before, such corrections can familiarize them with misinformation they were not yet familiar with. In other words, misinformation corrections that repeat the to-be-corrected false claims can disseminate misinformation to new audiences (Schwarz, Newman, & Leach, 2016). For example, if you have never heard the myth about vaccines

Figure 1. Example of an online fact-check, repeating the false claim and adding a 'false' tag. A link at the bottom provides access to additional information.

causing autism, encountering a correction of this falsehood signals that someone once believed, or still believes, that they do.

This view suggests that communicators should focus on factual information without repeating the misinformation, and in particular that simply restating misinformation with a 'false' tag should be avoided – memory for the tag could be lost and the misinformation may later be erroneously relied upon as it is easily retrieved or recognized as familiar (see Gilbert, Krull, & Malone, 1990; Mayo, Schul, & Burnstein, 2004). The first aim of the current study was to test whether simple false-tag fact-checks are indeed ineffective or potentially harmful.

Despite these potential issues, focusing exclusively on factual information would, of course, defeat the purpose of a fact-check (i.e., usually the false claim will have to be repeated in order to correct it), and thus, the question regarding the best fact-checking format remains. Empirically, it has been established that providing a detailed refutation is more effective than providing a simple 'X is not true' retraction (Chan, Jones, Hall Jamieson, & Albarracín, 2017; Ecker, Lewandowsky, & Tang, 2010; Johnson & Seifert, 1994; Nyhan & Reifler, 2015c; Seifert, 2002; Swire, Ecker, Hogan, & Lewandowsky, 2017; Walter & Murphy, 2018). A refutation can explain why the misinformation is false and can provide alternative factual information to replace the debunked misinformation in a person's mental model. Under such circumstances, it has been found that repeating misinformation when correcting it is actually beneficial, resulting in *reduced* post-correction misinformation reliance (Ecker *et al.*, 2017). This supports theoretical frameworks that suggest misinformation repetition can have a positive effect, by making a conflict salient between a factual account and an incorrect knowledge representation. The co-activation of corresponding correct and incorrect representations is thought to facilitate integration of new factual information into an existing, flawed mental model and can thus be considered a prerequisite for successful model updating and knowledge revision (Elsey & Kindt, 2017; Kendeou, Walsh, Smith, & O'Brien, 2014; Putnam, Wahlheim, & Jacoby, 2014).

Thus, the existing literature suggests that effective fact-checks should provide detailed refutations that repeat the misinformation in order to debunk it, while specifying both the reason for the misinformation being wrong and an alternative, factual account. However, one difficulty with this approach lies in the space constraints associated with social-media communications. Even though Twitter has now doubled its character allowance to 280, this is by no means ample space for a detailed refutation. The second aim of the present study was thus to test whether a refutational approach is more effective than plain retractions (i.e., false-tag fact-checks) even if the refutation is constructed under severe space constraints.

To this end, we collated a selection of true and false statements. True statements were simply affirmed – that is, repeated and labelled with a 'true' tag. False statements were corrected in one of two ways. One type of correction implemented a simple retraction – the false claim was restated and labelled with a 'false' tag. This retraction condition was contrasted with a refutation condition. Refutations were kept to below approximately 140 characters (the study was planned and designed before Twitter relaxed its character limit). In designing the refutations, we followed a few simple guidelines: The fact-check first highlighted why the false claim was false (e.g., because it was based on a false statistic; see Seifert, 2002); it specified and discredited the source of the misinformation (e.g., a tabloid as opposed to a reputable source; see Guillory & Geraci, 2013); it warned people before exposing them to misinformation, so readers would be cognitively on guard when processing the false claim and would need to retrospectively re-evaluate the information

(see Ecker *et al.*, 2010); it repeated the misinformation in order to refute it in a salient manner (see Ecker *et al.*, 2017); it refuted the misinformation with a factually correct statement (see Johnson & Seifert, 1994); it provided a credible fact source (Guillory & Geraci, 2013; Vraga & Bode, 2018); and it supported the factual statement with a graphical representation of relevant data (see Mason *et al.*, 2017; Nyhan & Reifler, 2018). Both the refutations and the retractions used a 'false' tag – in the refutations, this was placed at the top of the fact-check, such that it served as an additional warning that the claim about to be encountered was contested; in the retraction, it was placed at the end, after the false claim (as is often the case in facts vs. myths materials). We measured claim belief before and/or after the manipulation; additionally, inferential-reasoning questions were used post-manipulation to indirectly assess claim belief, as well as use of the alternative, factual information provided with the refutation. The impact of refutations on claim belief and inferential reasoning was then contrasted with the impact of the standard false-tag-only retractions.

Our first main hypothesis was that detailed refutations would be more effective than simple retractions. We specified that: (H1a) False-claim belief will be lower after refutations than retractions. (H1b) Inferential-reasoning scores will be lower, reflecting lower endorsement of the contested claim, after refutations than retractions.

The second main hypothesis was that plain retractions would elicit familiarity backfire. We took into account that retractions can potentially backfire relative to various baselines and thus specified the following three sub-hypotheses: (H2a) A retraction will backfire relative to the pre-correction baseline in the same sample of participants. (H2b) A retraction will backfire in participants not previously exposed to the false claim. This tests if a retraction will spread misinformation to new audiences; that is, participants who receive just a retraction will demonstrate greater post-correction belief in a false claim than participants upon initial exposure to the false claim. (H2c) Similarly, a retraction will backfire in the sense that receiving just a retraction will lead to greater inferential-reasoning scores compared to baseline scores from participants never exposed to the false claim in the experiment.

There were two additional, secondary hypotheses. One related to the general impact of claim familiarity, namely that repeated claim exposure may lead to greater claim belief. We specified: (H3) Three claim exposures (the initial presentation to measure pre-manipulation belief, the subsequent fact-check, and at test to measure post-manipulation belief) will be associated with greater claim belief at test relative to conditions involving only two exposures (in the fact-check and at test). The final hypothesis tested whether refutations could ironically reduce *fact* beliefs, as detailed refutations might make the simple affirmations less convincing in comparison. We specified: (H4) Belief in true claims will be lower in refutation compared to retraction conditions.

Method

We ran two online experiments. As the experiments were almost identical in design, differing in only two aspects, we will report them together. The experiments used true and false claims across various experimental conditions; claim veracity was manipulated within subjects, and experimental condition was a between-subjects factor. Experiment 1 had five conditions, thus implementing a 2×5 within–between design. Conditions were as follows: (1) In the retraction condition, participants were presented with true and false claims, rated their claim belief for each (time 1), and were then given affirmations and plain retractions repeating the claims and labelling them as true or false,

respectively. At test (time 2), participants rerated their claim beliefs and responded to a series of inferential-reasoning questions relating to the claims. (2) The refutation condition was identical, but instead of just retracting false claims as false, detailed refutations were provided. (3) The retraction-only and (4) refutation-only conditions were identical to the first two conditions, but participants were not exposed to the claims initially; that is, they received only the affirmations and retractions/refutations, and rated their claim beliefs only at time 2. (5) Participants in the no-exposure condition received only the inference questions and were thus never exposed to the claims or any corrections.[1] Experiment 2 had the same conditions except the no-exposure condition and thus had a 2 × 4 within–between design. The retention interval between fact-checks and test was approximately 1 day in Experiment 1 and approximately 1 week in Experiment 2.

Participants

Participants were U.S. residents recruited via Amazon MTurk. Participants in both experiments were randomly assigned to conditions (with the constraint of approximately equal cell numbers, $n \approx 125$ per condition; also, condition 5 of Experiment 1 was necessarily run separately from the other conditions, due to a significant difference in testing time/payment, and thus, condition assignment was not random for that condition). In Experiment 1, 518 participants completed the 'study' phase of conditions 1–4; of these, 441 participants also completed the test phase (retention rate approximately 85%). An additional 125 participants completed the single-phase condition 5. Of the resulting 566 participants, the data from 35 participants (6.2%) were excluded based on *a priori* exclusion criteria relating to completion time, systematically inconsistent or uniform responding, etc. (criteria and exclusion numbers are specified in the Appendix), resulting in a total sample size of $N = 531$. This comprised 280 women, 250 men, and one participant of undisclosed gender. Mean age was $M = 39.0$ years ($SD = 11.9$; range: 21–77 years).

In Experiment 2, a separate sample of 509 participants completed the study phase, of which 412 completed the test phase (retention rate approximately 81%). Based on the exclusion criteria, 43 data sets (10.4 %) were excluded, resulting in a final sample of $N = 369$. This comprised 199 women, 166 men, and four participants of undisclosed gender. Mean age was $M = 39.2$ years ($SD = 12.0$; range: 20–76 years).

Materials and procedure

We selected a range of 'fact-checkable' claims (Merpert, Furman, Anauati, Zommer, & Taylor, 2018) from various online and social-media sources. While some claims were apolitical, the remainder represented a range of views from conservative to liberal. Claims had fewer than 140 characters and were presented as social-media posts similar to 'tweets'; each post was associated with a fictional source handle (e.g., '@StacyFury'); a coloured, circular icon containing the first letter of the handle (e.g., 'S') was used instead of a profile picture (the icons were similar to the default icons used for Google accounts). Figure 2 shows an example true and false claim (all claims, as well as refutations and inference questions, are provided in the Appendix S1).

[1] *Research in this area often implements an additional control condition where claims are presented without fact-checking. This condition was not necessary to test our main hypotheses in this study and was thus omitted for pragmatic and cost reasons.*

In conditions 1 and 2, participants were first presented with all 12 claims (six true and six false) individually and in randomized order; participants rated their belief in each – whether they thought the claim was true or false – on a 0–10 scale. Participants in conditions 1–4 then received the affirmations and retractions/refutations (depending on condition; see Figure 3 for examples), again in randomized order. In the test phase, participants were presented with the claims (again) in randomized order and rated (conditions 3 and 4) or rerated (conditions 1 and 2) their belief in each claim. Additionally, all participants (including condition 5 of Experiment 1) responded to two inference questions per claim, where higher scores would indirectly reflect stronger belief in the claim. The first inference question involved an estimation of the true value relating to each claim (e.g., 'Out of 100 white murder victims in the US, how many do you think are murdered by black people? Enter a number between 0 and 100'.); the second inference question involved a rating of agreement/disagreement regarding a statement related to the respective claim, on a 0–10 scale (e.g., 'Please indicate how much you agree or disagree with the following statement: More resources are needed to deal with murders by black people'.). The two inference questions per claim were always grouped together and presented in the specified order (first, second).

The study had ethics approval from the university where the research was conducted. All participants initially read an approved information sheet and were provided a detailed debriefing sheet at the end of the experiment. The experiment was run via Qualtrics survey software (Qualtrics, Provo, UT, USA). The experiment took approximately 12 min. to complete (condition 5 of Experiment 1 only took approximately 3 min). Participants were reimbursed US$1.50 through Amazon MTurk (US$0.50 for completion of phase 1 and US$1 for completion of phase 2; participants in condition 5 of Experiment 1 received US$0.40). Testing was completed in May/June 2018.

Results

As a sanity check, we first tested whether there were any belief rating differences at time 1 (BR1) between retraction and refutation conditions (conditions 1 and 2), which were identical up to this point. In Experiment 1, a 2 × 2 analysis of variance (ANOVA) with the within-subjects factor claim veracity (true vs. false) and the between-subjects factor condition (retraction vs. refutation) yielded a significant main effect of claim veracity, $F(1, 203) = 84.78$, $MSE = 1.08$, $p < .001$, $\eta_p^2 = .29$, indicating that belief in true claims

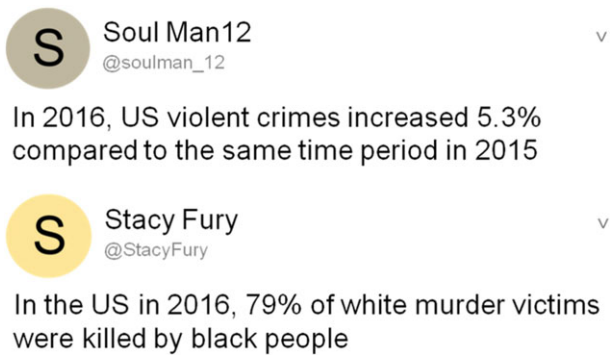

Figure 2. Example of true (top) and false (bottom) claim.

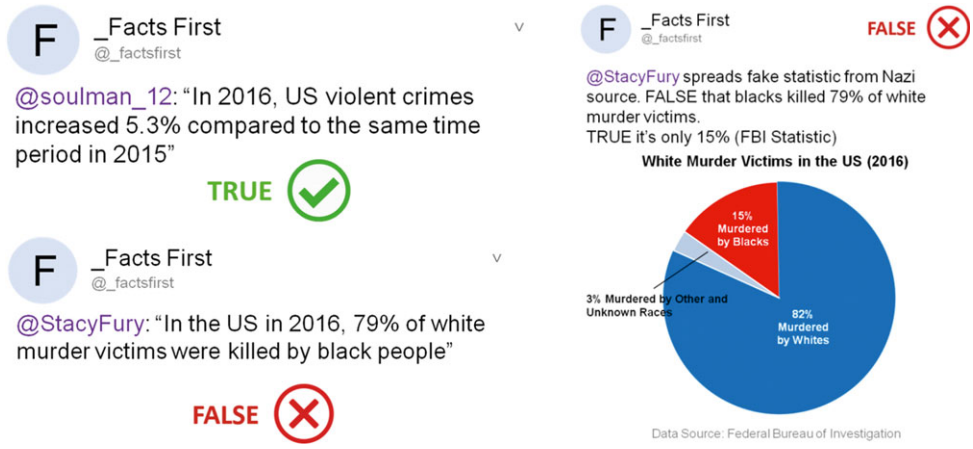

Figure 3. Example of affirmation (top left), plain retraction (bottom left), and refutation (right).

was significantly higher than belief in false claims (the main effect of condition and the interaction were non-significant, $F < 1$). Direct tests showed that neither initial belief in true nor initial belief in false claims differed across conditions 1 and 2, $M_{\text{BR1true-1}} = 5.63$ ($SE = .12$); $M_{\text{BR1true-2}} = 5.63$ ($SE = .12$); $M_{\text{BR1false-1}} = 4.66$ ($SE = .13$); $M_{\text{BR1false-2}} = 4.71$ ($SE = .14$); both $F < 1$.

In Experiment 2, the 2 × 2 ANOVA yielded a main effect of claim veracity, $F(1, 185) = 36.69$, $MSE = 1.19$, $p < .001$, $\eta_p^2 = .17$, again indicating that initial belief in true claims was significantly higher than initial belief in false claims. While the main effect of condition, $F(1, 185) = 2.20$, $MSE = 2.44$, $p = .14$, $\eta_p^2 = .01$, as well as the interaction, $F(1, 185) = 2.70$, $MSE = 1.19$, $p = .10$, $\eta_p^2 = .01$, was non-significant, direct tests showed that initial belief in false claims differed significantly across (identical) conditions 1 and 2, $M_{\text{BR1false-1}} = 5.34$ ($SE = .14$); $M_{\text{BR1false-2}} = 4.92$ ($SE = .14$); $F(1, 185) = 4.43$, $MSE = 1.91$, $p = .04$, $\eta_p^2 = .02$. Initial belief in true claims did not differ significantly between conditions, $M_{\text{BR1true-1}} = 5.84$ ($SE = .13$); $M_{\text{BR1true-2}} = 5.79$ ($SE = .14$).

H1a: False-claim belief will be lower after refutations than retractions.

Next, we addressed hypothesis H1a, namely whether false-claim belief would be lower after detailed refutations (conditions 2 and 4) than simple retractions (conditions 1 and 3).

In Experiment 1, the hypothesis was disconfirmed. False-claim belief ratings at time 2 (BR2) across the four conditions of interest are presented in Figure 4. While a one-way ANOVA yielded a main effect of condition, $F(3, 402) = 3.15$, $MSE = 3.97$, $p = .03$, $\eta_p^2 = .02$, planned direct contrasts showed no significant difference between conditions 1 and 2, $F < 1$, or between conditions 3 and 4, $F(1, 402) = 2.14$, $p = .14$, $\eta_p^2 = .01$. This indicates that refutations were not more effective than plain retractions at reducing claim belief after a 1-day retention interval. This was confirmed in an analysis of belief-change scores (BC = BR2 − BR1) across conditions 1 and 2, $F(1, 203) = 1.04$, $MSE = 4.12$, $p = .31$, $\eta_p^2 = .01$. While false-claim belief change was substantial, $M_{\text{BCfalse-1}} = -1.90$ ($SE = .20$); $M_{\text{BCfalse-2}} = -2.19$ ($SE = .20$), there was no difference between retraction and refutation conditions.

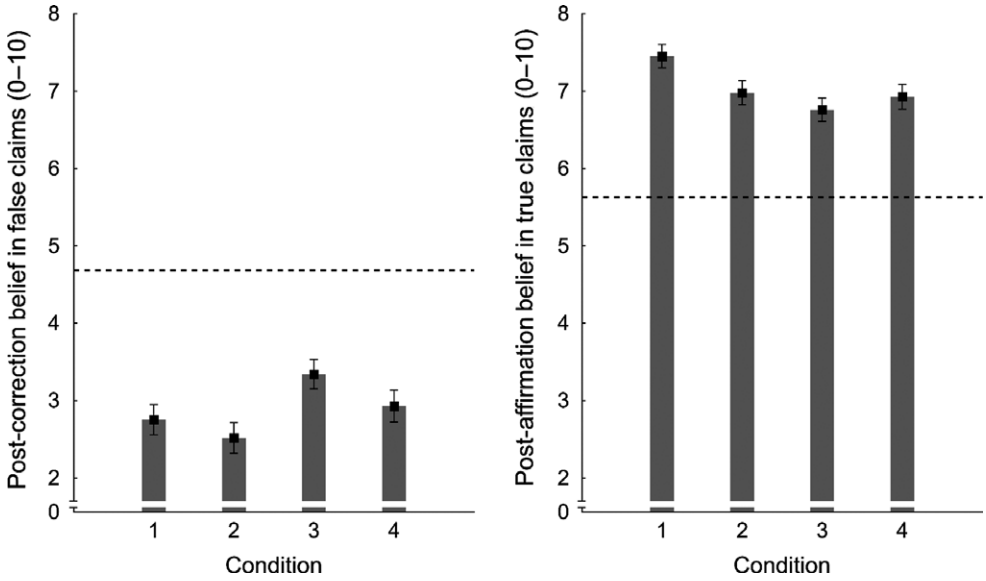

Figure 4. Mean post-correction/affirmation belief ratings regarding false (left panel) and true claims (right panel) across conditions in Experiment 1. Condition 1: retraction; condition 2: refutation; condition 3: retraction-only; and condition 4: refutation-only. Dotted lines indicate mean pre-correction/affirmation belief ratings from conditions 1 and 2.

In Experiment 2, however, with a longer retention interval, hypothesis H1a was supported. False-claim belief ratings at time 2 across conditions are presented in Figure 5. A one-way ANOVA yielded a main effect of condition, $F(3, 365) = 12.49$, $MSE = 3.20$, $p < .001$, $\eta_p^2 = .09$. Planned direct contrasts showed a significant difference between conditions 1 and 2, $F(1, 365) = 14.90$, $MSE = 3.20$, $p < .001$, $\eta_p^2 = .07$. There was no difference between conditions 3 and 4, $F < 1$. This indicates that refutations were more effective than plain false-tag retractions at reducing claim belief after a 1-week retention interval. This was confirmed in an analysis of belief-change scores across conditions 1 and 2 (which was important given the observed baseline difference in the belief rating at time 1), $F(1, 185) = 4.83$, $MSE = 3.30$, $p = .03$, $\eta_p^2 = .03$. False-claim belief change was greater after a refutation (condition 2) than after a retraction (condition 1), $M_{\text{BCfalse-1}} = -1.52$ ($SE = .19$); $M_{\text{BCfalse-2}} = -2.11$ ($SE = .19$).

H1b: False-claim inference scores will be lower after refutations than retractions.

Next, we tested hypothesis H1b, namely whether detailed refutations would lead to lower inferential-reasoning scores compared to plain retractions. This hypothesis was supported in both experiments. False-claim inference scores from all conditions in Experiments 1 and 2 are shown in Figures 6 and 7, respectively.

In Experiment 1, a one-way ANOVA yielded a main effect of condition, $F(4, 526) = 18.18$, $MSE = 1.08$, $p < .001$, $\eta_p^2 = .12$. Planned direct contrasts showed a significant difference between conditions 1 and 2, $F(1, 526) = 12.61$, $MSE = 1.08$, $p < .001$, $\eta_p^2 = .05$. Conditions 3 and 4 also differed significantly, $F(1, 526) = 5.30$, $MSE = 1.08$, $p = .02$, $\eta_p^2 = .02$.

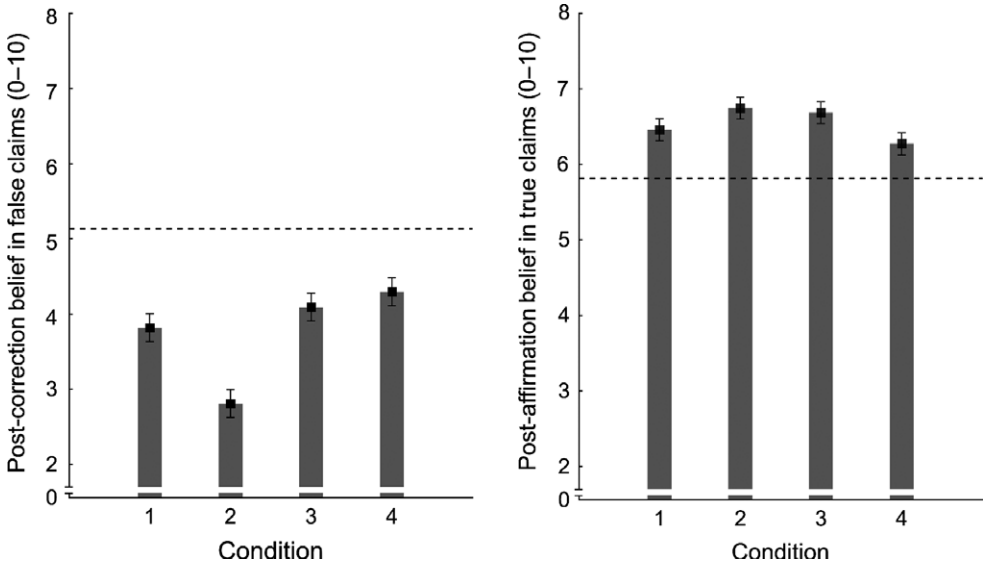

Figure 5. Mean post-correction/affirmation belief ratings regarding false (left panel) and true claims (right panel) across conditions in Experiment 2. Condition 1: retraction; condition 2: refutation; condition 3: retraction-only; and condition 4: refutation-only. Dotted lines indicate mean pre-correction/affirmation belief ratings from conditions 1 and 2.

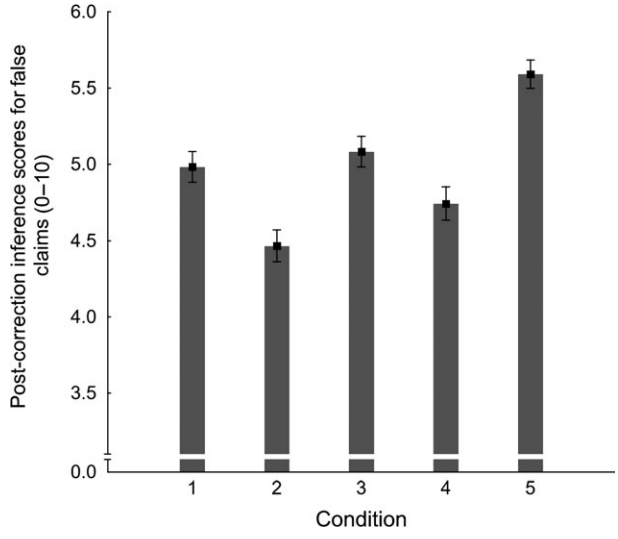

Figure 6. Mean post-correction inferential-reasoning scores regarding false claims across conditions in Experiment 1. Condition 1: retraction; condition 2: refutation; condition 3: retraction-only; condition 4: refutation-only; and condition 5: no-exposure.

In Experiment 2, a one-way ANOVA yielded a main effect of condition, $F(3, 365) = 4.00$, $MSE = 0.73$, $p = .008$, $\eta_p^2 = .03$. Planned direct contrasts showed a significant difference between conditions 1 and 2, $F(1, 365) = 8.30$, $MSE = 0.73$,

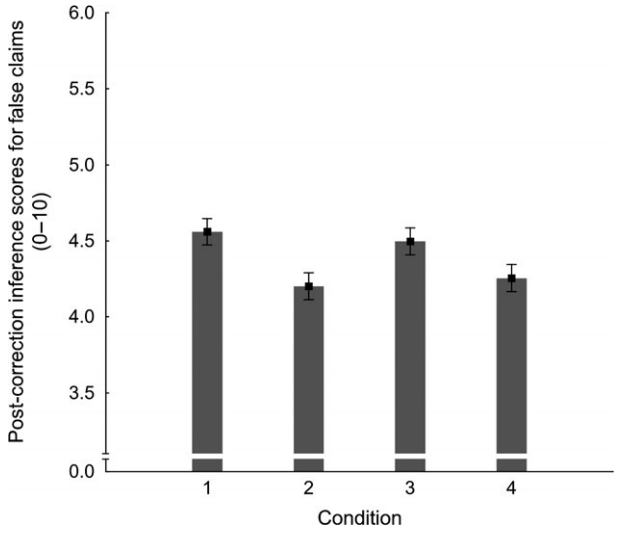

Figure 7. Mean post-correction inferential-reasoning scores regarding false claims across conditions in Experiment 2. Condition 1: retraction; condition 2: refutation; condition 3: retraction-only; and condition 4: refutation-only.

$p = .004$, $\eta_p^2 = .04$. The difference between conditions 3 and 4 was only marginally significant, $F(1, 365) = 3.70$, $MSE = 0.73$, $p = .06$, $\eta_p^2 = .02$.

H2a: A retraction may ironically increase false-claim belief from pre- to post-correction.

This hypothesis was not supported: Both plain retractions and refutations reduced belief in false claims, as evident in the consistently negative belief-change scores listed under hypothesis (H1a) earlier. The belief reduction associated with a plain retraction was substantial and significant in both Experiment 1, $F(1, 105) = 86.40$, $MSE = 2.22$, $p < .001$, $\eta_p^2 = .45$, and Experiment 2, $F(1, 94) = 61.58$, $MSE = 1.79$, $p < .001$, $\eta_p^2 = .40$.

H2b: False-claim belief may be greater after retraction than unchallenged exposure.

Next, we tested whether plain retractions might backfire in participants not previously exposed to the false claim – that is, participants who receive just a retraction might demonstrate greater post-correction belief in a false claim than participants upon initial exposure to the false claim. This is one way to test the idea that presenting retractions might cause harm by spreading misinformation to a new audience. This hypothesis was disconfirmed in both experiments. In Experiment 1, time-2 belief ratings in conditions 3 and 4 (see Figure 4) were substantially lower than time-1 belief ratings of conditions 1 and 2 (combined), $M_{\text{BR2false-3}} = 3.34$ ($SE = .19$); $M_{\text{BR2false-4}} = 2.93$ ($SE = .21$); $M_{\text{BR1false-1/2}} = 4.68$ ($SE = .13$). Likewise, in Experiment 2, time-2 belief ratings in conditions 3 and 4 (see Figure 5) were substantially lower than time-1 belief ratings of conditions 1 and 2 (combined), $M_{\text{BR2false-3}} = 4.09$ ($SE = .19$); $M_{\text{BR2false-4}} = 4.30$ ($SE = .19$); $M_{\text{BR1false-1/2}} = 5.13$ ($SE = .14$). In other words, retractions and refutations *always reduced* false-claim belief.

H2c: False-claim-congruent reasoning may be enhanced after retraction compared to no exposure.

Next, we tested whether plain retractions might backfire in the sense that receiving just a retraction might lead to greater inferential-reasoning scores compared to baseline scores from participants never exposed to the false claim in the experiment. This hypothesis could only be tested in Experiment 1; it was disconfirmed. Specifically, as can be seen in Figure 6, all conditions receiving refutations or retractions had lower inference scores than the no-exposure control group (condition 5), all $F(1, 526) > 13.79, p < .001$, $\eta_p^2 > .07$.

H3: Repeated false-claim exposure may lead to greater claim belief.

Next, we tested whether three exposures to false and true claims (in the initial presentation to measure pre-manipulation belief, the subsequent fact-check, and at test to measure post-manipulation belief) might be associated with greater claim belief at test based on increased claim familiarity, relative to conditions involving only two exposures (in the fact-check and at test). In other words, the additional repetition of myths and facts in conditions 1 and 2 might be associated with greater post-manipulation belief ratings. This hypothesis was disconfirmed in both experiments.

In Experiment 1, Figure 4 shows the opposite pattern for false claims: Additional repetition, if anything, was associated with decreased false-claim belief, with lower time-2 belief ratings in conditions 1 and 2 (combined) than in conditions 3 and 4 (combined), $M_{\text{BR2false-1/2}} = 2.64$ ($SE = .20$); $M_{\text{BR2false-3/4}} = 3.14$ ($SE = .20$); $F(1, 402) = 6.32$, $MSE = 3.97$, $p = .01$, $\eta_p^2 = .02$. In other words, refutations were especially powerful if participants were previously exposed to the false claim. Time-2 belief ratings for true claims are shown in Figure 4 (right panel). A one-way ANOVA yielded a main effect of condition, $F(3, 402) = 3.78$, $MSE = 2.44$, $p = .01$, $\eta_p^2 = .03$. Again, these data show a positive effect of repetition, with greater post-manipulation belief ratings in conditions 1 and 2 (combined) than in conditions 3 and 4 (combined), $M_{\text{BR2true-1/2}} = 7.22$ ($SE = .15$); $M_{\text{BR2true-3/4}} = 6.84$ ($SE = .16$); $F(1, 402) = 5.65$, $MSE = 2.44$, $p = .02$, $\eta_p^2 = .02$ (although this effect was due almost exclusively to the higher fact-belief ratings in condition 1). Altogether, this suggests additional repetition generally *increases* belief accuracy at test.

Experiment 2 likewise showed that additional repetition was associated with decreased false-claim belief (see Figure 5). Belief ratings at time 2 were lower in conditions 1 and 2 (combined) than in conditions 3 and 4 (combined), $M_{\text{BR2false-1/2}} = 3.31$ ($SE = .18$); $M_{\text{BR2false-3/4}} = 4.20$ ($SE = .19$); $F(1, 365) = 22.37$, $MSE = 3.20$, $p < .001$, $\eta_p^2 = .08$ (although this effect was due almost exclusively to the lower false-claim belief ratings in condition 2). In other words, refutations were especially powerful if participants were previously exposed to the false claim. Time-2 belief ratings for true claims are shown in Figure 5 (right panel). A one-way ANOVA yielded a non-significant main effect of condition, $F(3, 365) = 2.20$, $MSE = 1.94$, $p = .09$, $\eta_p^2 = .02$. There was no effect of repetition, with equivalent post-manipulation belief ratings in conditions 1 and 2 (combined) compared to conditions 3 and 4 (combined), $M_{\text{BR2true-1/2}} = 6.60$ ($SE = .14$); $M_{\text{BR2true-3/4}} = 6.48$ ($SE = .15$); $F < 1$.

H4: Refutations might reduce belief in true claims.

Finally, we tested whether refutations might reduce beliefs in *true* claims, as detailed refutations might make the simple affirmations of factual claims less convincing in comparison. The data from both experiments yielded some evidence in support of this hypothesis.

The data in Figure 4 (right panel) show that in Experiment 1, belief in true claims in conditions 2 and 4 (combined) was not significantly lower than factual beliefs in conditions 1 and 3 (combined), $F < 1$. However, we found that true-claim belief was greater in condition 1 than in condition 2, $F(1, 402) = 4.66$, $MSE = 2.44$, $p = .03$, $\eta_p^2 = .02$. For the sake of completeness, we also examined inferential-reasoning scores (IS) relating to true claims across conditions. Means were $M_{\text{IStrue-1}} = 5.60$ ($SE = .09$); $M_{\text{IStrue-2}} = 5.53$ ($SE = .09$); $M_{\text{IStrue-3}} = 5.53$ ($SE = .09$); $M_{\text{IStrue-4}} = 5.45$ ($SE = .10$); and $M_{\text{IStrue-5}} = 5.13$ ($SE = .08$). A one-way ANOVA yielded a main effect of condition, $F(4, 526) = 4.58$, $MSE = 0.88$, $p = .001$, $\eta_p^2 = .03$. The fact-related inference score was lower in the no-exposure group (condition 5) compared to all other conditions, all $F(1, 526) > 6.09$, $p < .01$, $\eta_p^2 > .03$. Conditions 1–4 did not differ from each other, all $F < 1.25$.

The data in Figure 5 (right panel) show that in Experiment 2, belief in true claims in conditions 2 and 4 (combined) was not significantly lower than factual beliefs in conditions 1 and 3 (combined), $F < 1$. However, we found that true-claim belief was greater in condition 3 than in condition 4, $F(1, 365) = 3.95$, $MSE = 1.94$, $p \leq .05$, $\eta_p^2 = .02$. For the sake of completeness, we also examined inferential-reasoning scores (IS) relating to true claims across conditions. Means were $M_{\text{IStrue-1}} = 4.71$ ($SE = .09$); $M_{\text{IStrue-2}} = 4.70$ ($SE = .09$); $M_{\text{IStrue-3}} = 4.89$ ($SE = .09$); and $M_{\text{IStrue-4}} = 4.71$ ($SE = .09$). A one-way ANOVA yielded no main effect of condition, $F(3, 365) = 1.02$, $MSE = 0.75$, $p = .38$, $\eta_p^2 = .01$.

Discussion

The current research investigated the effectiveness of online fact-checks, with an experimental paradigm that attempted to mimic to some extent the real-world environment of social media. Before discussing the results, we first acknowledge some limitations. First, it is known that online convenience samples are not fully representative of the general population. Thus, the participants may be more familiar with, and perhaps more open to, online fact-checking. This means that the general effectiveness of fact-checking may be overestimated in this research. Secondly, in order to investigate the impact of the refutational format and prior exposure to fact-checked claims independently of source credibility, we elected to use fictional social-media accounts with plain icons. This differs from the real world, of course, where social-media consumers can more readily ascribe characteristics to information sources, such as trustworthiness and perceived expertise. This means that in the real world, misinformation may have a stronger impact if it comes from a familiar source that is perceived as trustworthy, but also that the effectiveness of fact-checks may be greater if the fact-checks come from well-known, trusted fact-checkers (e.g., see Guillory & Geraci, 2013; Margolin *et al.*, 2018; Swire, Berinsky *et al.*, 2017), which might imply that the general effectiveness of fact-checking may be *under*estimated in this research. Exploring the impact of source credibility was beyond the scope of the present work, although we note that source credibility should have no impact on the assessment of differences between our experimental conditions.

The present study aimed to answer two questions: (1) Are fact-checks of a particularly common format – those that repeat a false claim and simply tag it as false – ineffective and potentially harmful? (2) Are refutations that implement a number of best-practice recommendations, but under severe space limitations, more effective than such false-tag-only retractions at reducing belief in false claims and false-claim-congruent reasoning?

Question (1) relates to the notion that corrections may inadvertently backfire when they boost misinformation familiarity through repetition, without providing details regarding the reasons for the assessment of the claim as false, or alternative factual information (Lewandowsky *et al.*, 2012; Peter & Koch, 2016; Schwarz *et al.*, 2016; Skurnik *et al.*, 2005). More specifically, a correction could theoretically backfire with reference to two separate baselines: Post-correction belief in a false claim could be higher than (1) pre-correction belief in the same sample or (2) claim belief in a different sample. While a within-subjects contrast provides a more conservative test of a familiarity backfire effect, it may be affected more strongly by demand characteristics; a contrast with a different sample introduces between-sample variability but is better suited to capture the notion that corrections may backfire by disseminating misinformation to new audiences, viz. people who have not encountered the false claim before. The present study allowed us to test both possibilities. The results were clear-cut and showed that simple retractions – messages that repeated a false claim while tagging it as false – did not backfire relative to either baseline. In fact, retractions substantially reduced belief in false claims relative to the pre-correction level in the same sample, as well as relative to the level of belief expressed by a different sample after initial, unchallenged exposure to the false claims. This pattern was consistent across both experiments. Moreover, Experiment 1, with its inclusion of a condition in which participants were never exposed to the claims, showed that false-claim-congruent inferential reasoning was reduced by a retraction. Thus, the present study provides no support for the existence of familiarity backfire effects (in line with Swire, Ecker *et al.*, 2017)[2]; there does not seem to be any harm associated with simple false-tag fact-checks. Moreover, additional claim repetition was generally associated with enhanced accuracy at test: Across both experiments, refutations were more powerful if participants were previously exposed to the myth, which runs counter to the assumption that greater familiarity with claims drives greater endorsement. However, there is one additional limitation that prevents us from concluding that familiarity backfire effects do not exist: While we aimed to use reasonably obscure claims, we have no way of ruling out that some of the claims were at least somewhat familiar to some participants. Thus, we cannot rule out the existence of a familiarity backfire effect relative to a no-exposure baseline with entirely novel claims. Future research should therefore investigate this possibility.

Of course, not causing any harm is the lowest possible demand one should place on fact-checking. Question (2) thus focused on the notion that the impact of a fact-check can potentially be increased by following some simple refutation guidelines, even under severe space constraints. Evidence for this notion was mixed. In Experiment 1, with a retention interval of approximately 1 day, simple retractions were as effective as refutations at reducing belief in false claims. This may reflect the integrity of participants' memory for the simple fact-checks – that is, which claims were affirmed and which retracted – after a relatively brief delay, such that there was no additional, significant

[2] *Please note that the familiarity backfire effect should be differentiated from the worldview backfire effect, for which there is also inconsistent evidence (see Ecker & Ang, 2018; Lewandowsky* et al.*, 2012; Nyhan & Reifler, 2010; Wood & Porter, 2018).*

benefit associated with the refutational format. However, Experiment 2, with a retention interval of approximately 1 week, yielded evidence in favour of the refutational format. In other words, it seems that detailed refutations are associated with a more sustained reduction in false beliefs. After a week, recollection for which claims were affirmed and which retracted will have faded (while claim familiarity may still be high), such that some retracted claims may again be accepted as valid (also see Swire, Ecker *et al.*, 2017; Swire, Berinsky, *et al.*, 2017). Indeed, the effectiveness of mere retractions – still the most common fact-checking format – after a 1-week delay was worryingly low. Refutations, on the other hand, naturally provide more recollectable details; such recollections may protect against false acceptance and thus support the accurate appraisal of false claims after a delay (see Ecker, Lewandowsky, Swire, & Chang, 2011; Seifert, 2002). Alternatively, refutations may lead to immediate belief reduction arising from factors that may make this change more sustained, despite the change not being quantitatively greater than the retraction-induced change initially. The change-driving factors may include enhanced scepticism towards the misinformation source (Lewandowsky, Stritzke, Oberauer, & Morales, 2005) or the highlighting of inconsistencies between the factual details provided and the inaccurate belief (Kendeou *et al.*, 2014; for a similar finding and interpretation, see Swire, Ecker *et al.*, 2017). We also argue that the longer retention interval in Experiment 2 and the associated forgetting of correction/affirmation details explain why the difference between true and false-claim belief ratings (post-correction/affirmation) was generally smaller in Experiment 2 compared to Experiment 1.

Moreover, across both experiments, the refutational format reduced misinformation-consistent reasoning. This means that the provision of additional factual information in the refutation allowed participants to be more in tune with the relevant true state of affairs and arrive at more evidence-based opinions. We can thus conclude that embedding a rebuttal in a fact-oriented context has beneficial implications beyond specific belief reduction, fostering a more sceptical and evidence-based approach to the issue at hand. This meshes well with educational literature that has shown that refutational approaches to teaching outperform traditional fact-based teaching (Cook, Bedford, & Mandia, 2014; Kowalski & Taylor, 2009).

As mentioned earlier, the impact of refutations seemed especially strong if participants were previously exposed to the false claim: Across all measures (except the post-affirmation fact-belief ratings in Experiment 1), the best outcome was associated with initial exposure to the false information, followed by a detailed refutation (i.e., condition 2). Initial exposure to misinformation being associated with a better outcome may point to one potential weakness of the short-format refutation: Despite the false claim being repeated in the refutation, the fact that much information is crammed into very little space may reduce the clarity and salience of the communication in participants with no prior representation of the challenged claim. Future research should investigate whether more prominently repeating the false claim in order to refute it may further improve corrective impact, for example by combining a false-tag retraction with a short-format refutation that is brief enough to be communicated directly, without the need to actively seek out additional information (e.g., via a hyperlink to a separate website) – although of course such additional information is likely to provide additional benefits for readers with the resources and motivation to peruse it.

Finally, in the present study, true claims were always just affirmed briefly with a 'true' tag. We speculated that this may make the affirmations less convincing and/or memorable when paired with detailed refutations. We found some support for this hypothesis, as refutation conditions (more specifically, the refutation condition in Experiment 1 and the

refutation-only condition in Experiment 2) tended to be associated with somewhat lower fact-belief ratings at time 2. While these effects were small and not entirely consistent across experiments, the pattern does suggest tentatively that fact-checkers should devote the same attention to true and false claims; that is, facts should also be affirmed by providing detailed affirmations – to strengthen the affirmations not just in absolute but also in relative terms.

Acknowledgments

This research was supported by grant DP160103596 from the Australian Research Council to the first author. The laboratory website can be found at www.emc-lab.org.

References

Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism and Mass Communication Quarterly*, *95*, 28–48. https://doi.org/10.1177/1077699016678186

Berinsky, A. J. (2015). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*, 241–262. https://doi.org/10.1017/s0007123415000186

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*, 619–638. https://doi.org/10.1111/jcom.12166

Brandtzaeg, P. B., & Følstad, A. (2017). Trust and distrust in online fact-checking services. *Communications of the ACM*, *60*, 65–71. https://doi.org/10.1145/3122803

Chan, M.-P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*, 1531–1546. https://doi.org/10.1177/0956797617714579

Cobb, M. D., Nyhan, B., & Reifler, J. (2013). Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology*, *34*, 307–326. https://doi.org/10.1111/j.1467-9221.2012.00935.x

Cook, J., Bedford, D., & Mandia, S. (2014). Raising climate literacy through addressing misinformation: Case studies in agnotology-based learning. *Journal of Geoscience Education*, *62*, 296–306. https://doi.org/10.5408/13-071.1

Dechêne, A., Stahl, C., Hansen, J., & Wanke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*, 238–257. https://doi.org/10.1177/1088868309352251

Ecker, U. K. H., & Ang, L. C. (2018). Political attitudes and the processing of misinformation corrections. *Political Psychology.* https://doi.org/10.1111/pops.12494

Ecker, U. K. H., Hogan, J. L., & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*, 185–192. https://doi.org/10.1016/j.jarmac.2017.01.014

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin and Review*, *18*, 570–578. https://doi.org/10.3758/s13423-011-0065-1

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory and Cognition*, *38*, 1087–1100. https://doi.org/10.3758/MC.38.8.1087

Elsey, J. W., & Kindt, M. (2017). Tackling maladaptive memories through reconsolidation: From neural to clinical science. *Neurobiology of Learning and Memory*, *142*, 108–117. https://doi.org/10.1016/j.nlm.2017.03.007

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*, 127–151. https://doi.org/10.1080/10584609.2014.914613

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*, 617–637. https://doi.org/10.1111/jcom.12038

Gilbert, D. T., Krull, D., & Malone, P. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, *59*, 601–613. https://doi.org/10.1037/0022-3514.59.4.601

Gottfried, J. A., Hardy, B. W., Winneg, K. M., & Jamieson, K. H. (2013). Did fact checking matter in the 2012 presidential campaign? *American Behavioral Scientist*, *57*, 1558–1567. https://doi.org/10.1177/0002764213489012

Graves, L. (2017). Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture and Critique*, *10*, 518–537. https://doi.org/10.1111/cccr.12163

Guillory, J. J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition*, *2*, 201–209. https://doi.org/10.1016/j.jarmac.2013.10.001

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Kendeou, P., Walsh, E. K., Smith, E. R., & O'Brien, E. J. (2014). Knowledge revision processes in refutation texts. *Discourse Processes*, *51*, 374–397. https://doi.org/10.1080/0163853X.2014.913961

Kowalski, P., & Taylor, A. K. (2009). The effect of refuting misconceptions in the introductory psychology class. *Teaching of Psychology*, *36*, 153–159. https://doi.org/10.1080/00986280902959986

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., . . . Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science*, *359*, 1094–1096. https://doi.org/10.1126/science.aao2998

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*, 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

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*, 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Stritzke, W. G. K., Oberauer, K., & Morales, M. (2005). Memory for fact, fiction, and misinformation. *Psychological Science*, *16*, 190–195. https://doi.org/10.1111/j.0956-7976.2005.00802.x

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, *35*, 196–219. https://doi.org/10.1080/10584609.2017.1334018

Mason, L., Baldi, R., Di Ronco, S., Scrimin, S., Danielson, R. W., & Sinatra, G. M. (2017). Textual and graphical refutations: Effects on conceptual change learning. *Contemporary Educational Psychology*, *49*, 275–288. https://doi.org/10.1016/j.cedpsych.2017.03.007

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*, 433–449. https://doi.org/10.1016/j.jesp.2003.07.008

Merpert, A., Furman, M., Anauati, M. V., Zommer, L., & Taylor, I. (2018). Is that even checkable? An experimental study in identifying checkable statements in political discourse. *Communication Research Reports*, *35*, 48–57. https://doi.org/10.1080/08824096.2017.1366303

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., & Reifler, J. (2015a). *Estimating fact-checking's effects: Evidence from a long-term experiment during campaign 2014.* Unpublished manuscript. Retrieved from http://www.americanpressinstitute.org/wp-content/uploads/2015/04/Estimating-Fact-Checkings-Effect.pdf

Nyhan, B., & Reifler, J. (2015b). The effect of fact-checking on elites: A field experiment on U.S. state legislators. *American Journal of Political Science*, *59*, 628–640. https://doi.org/10.1111/ajps.12162

Nyhan, B., & Reifler, J. (2015c). Displacing misinformation about events: An experimental test of causal corrections. *Journal of Experimental Political Science*, *2*, 81–93. https://doi.org/10.1017/XPS.2014.22

Nyhan, B., & Reifler, J. (2018). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*. https://doi.org/10.1080/17457289.2018.1465061

Peter, C., & Koch, T. (2016). When debunking scientific myths fails (and when it does not): The backfire effect in the context of journalistic coverage and immediate judgments as prevention strategy. *Science Communication*, *38*, 3–25. https://doi.org/10.1177/1075547015613523

Putnam, A. L., Wahlheim, C. N., & Jacoby, L. L. (2014). Memory for flip-flopping: Detection and recollection of political contradictions. *Memory and Cognition*, *42*, 1198–1210. https://doi.org/10.3758/s13421-014-0419-9

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick and the myths fade: Lessons from cognitive psychology. *Behavioural Science and Policy*, *2*, 85–95. https://doi.org/10.1353/bsp.2016.0009

Seifert, C. M. (2002). The continued influence of misinformation in memory: What makes a correction effective? *The Psychology of Learning and Motivation*, *41*, 265–292. https://doi.org/10.1016/S0079-7421(02)80009-3

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLoS ONE*, *13*, e0196087. https://doi.org/10.1371/journal.pone.0196087

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*, 233–255. https://doi.org/10.1111/jcom.12284

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*, 713–724. https://doi.org/10.1086/426605

Stencel, M. (2015). *'Fact check this': How U.S. politics adapts to media scrutiny.* Retrieved from https://www.americanpressinstitute.org/fact-checking-project/fact-checking-research/u-s-politics-adapts-media-scrutiny/single-page/

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science*, *4*, 160802. https://doi.org/10.1098/rsos.160802

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*, 1948–1961. https://doi.org/10.1037/xlm0000422

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, *33*, 460–480. https://doi.org/10.1080/10584609.2015.1102187

Vraga, E. K., & Bode, L. (2018). I do not believe you: How providing a source corrects health misperceptions across social media platforms. *Information, Communication and Society*, *21*, 1337–1353. https://doi.org/10.1080/1369118X.2017.1313883

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

Weaver, K., Garcia, S. M., Schwarz, N., & Miller, D. T. (2007). Inferring the popularity of an opinion from its familiarity: A repetitive voice can sound like a chorus. *Journal of Personality and Social Psychology*, *92*, 821–833. https://doi.org/10.1037/0022-3514.92.5.821

Wintersieck, A. L. (2017). Debating the truth: The impact of fact-checking during electoral debates. *American Politics Research*, *45*, 304–331. https://doi.org/10.1177/1532673X16686555

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

Wood, T., & Porter, E. (2018). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*. https://doi.org/10.1007/s11109-018-9443-y

Young, D. G., Jamieson, K. H., Poulsen, S., & Goldring, A. (2018). Fact-checking effectiveness as a function of format and tone: Evaluating FactCheck.org and FlackCheck.org. *Journalism and Mass Communication Quarterly*, *95*, 49–75. https://doi.org/10.1177/1077699017710453

Received 13 October 2018; revised version received 21 January 2019

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Exclusion criteria.

Appendix:

Exclusion criteria

The following exclusion criteria were set (number of excluded participants in Experiments 1/2 is provided in parentheses; note that some participants met multiple criteria, and thus, the total number of exclusions was lower than the summed numbers provided below).

1. Participants rating their English as 'poor' were excluded (n = 0/0).
2. Minimum completion times for study and test phases were set based on pilot testing (study phase: conditions 1 and 4: 120s; condition 2: 150s; condition 3: 90s; n = 6/6; test phase: conditions 1–4: 150s; condition 5 of Experiment 1: 120s; n = 8/3).
3. Participants were excluded if they selected the 'No, I did not put in a reasonable effort' response option to the question 'Should we use your data?' at the end of experiment (n = 1/1).
4. A minimal-variance criterion was set to identify uniform responding on rating items; based on pilot data, this was set as mean $SD < 0.5$ (on 0–10 response scale) across all time-2 belief ratings and second inference questions (n = 5/5);
5. Variance-outlier criterion was set to identify erratic responding, based on the same items as criterion (4); outliers were identified with the inter-quartile-range outlier-labelling rule, applying a 2.2 multiplier (n = 0/3).
6. To identify incoherent responding, two criteria were applied. (6) The first criterion targeted participants who systematically expressed strong (weak) belief, to then

estimate the true value as far from (close to) the claim value (e.g., stating 'I strongly believe it is true that 79% of white murder victims are killed by black people', and 'the true percentage of white murder victims killed by black people is 0%', or vice versa, 'I do not believe that it is true that 79% of white murder victims are killed by black people', and 'the true percentage of white murder victims killed by black people is 79%'). For each claim, we calculated the squared sum of belief rating 2 and the deviance from the claim value expressed in inference question 1, divided by the maximally possible value for the given claim, and excluded outliers on the mean of this score across all claims (again applying the outlier-labelling rule with a 2.2 multiplier; n = 7/11).

7. The second criterion was based on substantial positive mean item-wise correlations between belief rating 2 and the deviance from the claim value expressed in inference question 1. This correlation was negative for ~90% of participants – the more you believe a claim, the closer to the claimed value you estimate the true value to lie. Criterion was set at mean $r > .20$ (*a priori* criterion was 'positive,' but this would have excluded 44 participants; statistical significance was only reached at extreme values of $r \geq .59$ due to the small number of claims; n = 18/22).
8. While not specified *a priori*, also excluded were participants who completed the study phase twice (n = 0/2).