Journal of Applied Research in Memory and Cognition 9 (2020) 286–299

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

Review

Searching for the Backfire Effect: Measurement and Design Considerations

Briony Swire-Thompson*
Network Science Institute, Northeastern University, Boston, USA
Institute of Quantitative Social Science, Harvard University, Cambridge, USA

Joseph DeGutis
Boston Attention and Learning Laboratory, VA Boston Healthcare System, Boston, MA, USA
Department of Psychiatry, Harvard Medical School, Boston, MA, USA

David Lazer
Network Science Institute, Northeastern University, Boston, USA
Institute of Quantitative Social Science, Harvard University, Cambridge, USA

One of the most concerning notions for science communicators, fact-checkers, and advocates of truth, is the backfire effect; this is when a correction leads to an individual *increasing* their belief in the very misconception the correction is aiming to rectify. There is currently a debate in the literature as to whether backfire effects exist at all, as recent studies have failed to find the phenomenon, even under theoretically favorable conditions. In this review, we summarize the current state of the worldview and familiarity backfire effect literatures. We subsequently examine barriers to measuring the backfire phenomenon, discuss approaches to improving measurement and design, and conclude with recommendations for fact-checkers. We suggest that backfire effects are not a robust empirical phenomenon, and more reliable measures, powerful designs, and stronger links between experimental design and theory could greatly help move the field ahead.

General Audience Summary

A backfire effect is when people report believing even *more* in misinformation after they have seen an evidence-based correction aiming to rectify it. This review discusses the current state of the backfire literature, examines barriers to measuring this phenomenon, and concludes with recommendations for fact-checkers. Two backfire effects have gained popularity in the literature: the worldview backfire effect and the familiarity backfire effect. While these both result in increased belief after a correction, they occur due to different psychological mechanisms. The worldview backfire effect is said to occur when a person is motivated to defend their worldview because a correction challenges a person's belief system. In contrast, the familiarity backfire effect is presumed to occur when misinformation is repeated within the retraction. Failures to find or replicate both backfire effects have been widespread. Much of the literature has interpreted these failures to replicate to indicate that either (a) the backfire effect is difficult to elicit on the larger group level, (b) it is extremely item-, situation-, or individual-specific, or (c) the phenomenon does not exist at all. We suggest that backfire effects are not a robust empirical phenomenon, and that improved measures, more powerful designs, and stronger links between experimental design and theory, could greatly help move the field ahead. Fact-checkers can rest assured that it is *extremely unlikely* that their fact-checks will lead to increased belief at the group level.

Author Note

[*] Correspondence concerning this article should be addressed to Briony Swire-Thompson, Network Science Institute, Northeastern University, Boston, USA.

Furthermore, research has failed to show backfire effects systematically in the same subgroup, so practitioners should not avoid giving corrections to any specific subgroup of people. Finally, avoiding the repetition of the original misconception within the correction appears to be unnecessary and could even hinder corrective efforts. However, misinformation should always be *clearly and saliently* paired with the corrective element, and needless repetitions of the misconceptions should still be avoided.

Keywords: Backfire effects, Belief updating, Misinformation, Continued influence effect, Reliability

One of the most concerning notions for science communicators, fact-checkers, and advocates of truth is the *backfire effect*. A backfire effect occurs when an evidence-based correction is presented to an individual and they report believing even *more* in the very misconception the correction is aiming to rectify (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). This phenomenon has extremely important practical applications for fact-checking, social media, and all corrective communication efforts. However, there is currently a debate in the literature as to whether backfire effects exist at all, as recent studies have failed to find them, even under theoretically favorable conditions (e.g., Swire, Ecker, & Lewandowsky, 2017; Wood & Porter, 2019). In this article, we discuss the current state of the worldview and familiarity backfire effect literatures, examine barriers to measuring the correction of misinformation, and conclude with recommendations for fact-checkers and communicators.

Definitions

There are numerous barriers to changing inaccurate beliefs after corrections have been presented. The *continued influence effect* is where individuals still use inaccurate information in their reasoning and memory after a credible correction has been presented (Johnson & Seifert, 1994; Lewandowsky et al., 2012). There is also *belief regression*, where individuals initially update their belief after being exposed to the correction, but this belief change is not sustained over time (Berinsky, 2017; Kowalski & Taylor, 2017; Swire, Ecker et al., 2017). In contrast to the backfire effect, these barriers are where people at least still update their beliefs in the intended direction promoted by the correction. The term backfire effect only pertains to cases where a correction inadvertently *increases* misinformation belief relative to a precorrection or no-correction baseline. It has also been referred to as the boomerang effect (Hart & Nisbet, 2012) or backlash (Guess & Coppock, 2018).

Two backfire effects have gained popularity in the literature: the worldview backfire effect and the familiarity backfire effect. These both result in increased belief after a correction yet are thought to have different psychological mechanisms. The worldview backfire effect is said to occur when people are motivated to defend their worldview because a correction challenges their belief system (Cook & Lewandowsky, 2012). It is more likely to occur with items that are important to the individual, such as politicized "hot-button" issues or information that the individual believes in strongly (Flynn, Nyhan, & Reifler, 2017; Lewandowsky et al., 2012). In contrast to the mechanisms of the worldview backfire effect, the familiarity backfire effect is presumed to occur when misinformation is repeated within the correction (Schwarz, Sanna, Skurnik, & Yoon, 2007).[1] For example, if one were to try to correct a misconception and stated that "eating apricot seeds does NOT cure cancer," the correction repeats both "apricot seeds" and "curing cancer," thus making the original misinformation more familiar. This increased familiarity is problematic because people are more likely to assume that familiar information is true—a phenomenon called the *illusory truth effect* (Begg, Anas, & Farinacci, 1992). In other words, this boost in familiarity when correcting misinformation is thought to be sufficient to increase the acceptance of the misinformation as true, even though it is paired with a retraction.

Worldview Backfire Effect

The logic behind the worldview backfire effect stems from the motivated reasoning literature, where one's ideology and values influence how information is processed (Kunda, 1990; Wells, Reedy, Gastil, & Lee, 2009), and information that counters preexisting beliefs is evaluated more critically than belief-congruent information (Taber & Lodge, 2006). A possible reason for the backfire effect is that people generate counter-arguments consistent with their pre-existing views to contradict the new information or correction (Nyhan & Reifler, 2010).

The landmark paper regarding the worldview backfire effect is Nyhan and Reifler (2010). Their first experiment corrected the misconception that weapons of mass destruction were found in Iraq during the 2003 invasion. Liberal individuals, whose worldview aligned with the correction, were able to successfully update their belief, whereas conservatives increased their belief in the misconception. Although Nyhan and Reifler's second experiment failed to replicate the backfire effect for this item with the conservative group as a whole, they did find the phenomenon in a subset of conservative respondents who rated Iraq as the most *important* problem facing the country at that point in time. The authors suggested that backfire effects may only occur when a belief is strong, and the issue is currently connected with an individual's political identity. This suggestion aligns well with subsequent research demonstrating that worldview backfire effects have almost exclusively been found in either political or attitudinal subgroups, rather than communi-

[1] There are several studies that suggest that people misremember false information to be true more often than they misremember true information to be false (Peter & Koch, 2016; Skurnik, Yoon, Park, & Schwarz, 2005). Although this asymmetry could indeed stem from a familiarity process (see Swire, Ecker et al., 2017), this does not meet the criteria of a backfire effect. See Appendix A for details regarding articles that are frequently cited in support of backfire effects but do not meet backfire criteria.

ties as a whole. One major problem is that beyond the scientific literature, the media and online science blogs have often overgeneralized backfire effects found in subgroups to the population as a whole and to all corrective information (e.g., Science, 2017).

There have subsequently been worldview backfire effects reported in a variety of subgroups with misinformation regarding vaccines (in respondents with least favorable vaccine attitudes, Nyhan, Reifler, Richey, & Freed, 2014; in respondents with high levels of concern about vaccine side effects, Nyhan & Reifler, 2015), climate change (in Republican participants, Hart & Nisbet, 2012; in Republicans with high political interest, Zhou, 2016), the existence of death panels (in politically knowledgeable Palin supporters, Nyhan, Reifler, & Ubel, 2013), and with a fictitious scenario detailing that right-wing politicians generally misappropriate public funds more than left-wing politicians (in right-wing attentive participants, Ecker & Ang, 2019), see Appendix B. In addition to observing backfire effects in widely varying subgroups, a further complication is that the dependent variable has also varied substantially between studies. These dependent variables roughly fall into three categories: belief in or agreement with a claim (e.g., Nyhan & Reifler, 2010), behavioral intentions (e.g., Nyhan and Reifler, 2015), or use of misinformation when answering inference questions (e.g., Ecker & Ang, 2019).

Regardless of the dependent variable used, failures to find or replicate previously observed backfire effects have been widespread (e.g., Garrett, Nisbet, & Lynch, 2013; Nyhan, Porter, Reifler, & Wood, 2019; Schmid & Betsch, 2019; Swire, Berinsky, Lewandowsky, & Ecker, 2017; Swire-Thompson, Ecker, Lewandowsky, & Berinsky, 2019; Weeks, 2015; Weeks & Garrett, 2014), even when using identical items that previously elicited the phenomenon. For example, Haglin (2017) used identical methods and vaccine-related items to those from Nyhan and Reifler (2015) and failed to find any evidence of a backfire effect. The largest failure to replicate to-date was by Wood and Porter (2019), conducting five experiments with over 10,000 participants. The items were specifically chosen to be important ideological issues that would be theoretically conducive to a worldview backfire effect. The authors found that out of 52 issues corrected, no items triggered a backfire effect. Much of the literature has interpreted these failures to replicate to indicate that either (a) the backfire effect is difficult to elicit on the larger group level, (b) it is extremely item-, situation-, or individual-specific, or (c) the phenomenon does not exist at all. See Appendix B for details regarding which studies found a worldview backfire effect, which did not, and the dependent variable(s) used in each.

Familiarity Backfire Effect

In contrast to the ideological mechanisms behind the worldview backfire effect, familiarity backfire effects are often presumed to occur due to the correction increasing the misinformation's processing fluency. In other words, the correction of "apricot seeds do NOT cure cancer" increases the ease in which "apricot seeds" and "cancer" are retrieved and processed (Schwarz, Newman, & Leach, 2016). However, the specific mechanisms of *how* repetition could lead to an increase in perceived truth are currently under debate (see Unkelbach, Koch, Silva, & Garcia-Marques, 2019, for a review). Furthermore, the familiarity backfire effect has often been conflated with the more well-established illusory truth effect. The former refers to increasing belief due to information repetition *within* a correction and has little to no empirical support, whereas the latter refers to increasing belief due to information repetition in the *absence* of a correction and is a robust empirical phenomenon (Fazio, Brashier, Payne, & Marsh, 2015).

The original notion of the familiarity backfire effect stems from an unpublished manuscript (Skurnik, Yoon, & Schwarz, 2007), as cited in Schwarz et al., 2007) where participants who viewed a flyer with "myth vs. fact" information regarding the flu vaccine reported less favorable attitudes toward vaccination than those who did not view the flyer. Although this paper is highly cited (e.g., Berinsky, 2017; Cook, Bedford, & Mandia, 2014; Gemberling & Cramer, 2014; Lilienfeld, Marshall, Todd, & Shane, 2014; Peter & Koch, 2016; Pluviano, Watt, Ragazzini, & Della Sala, 2019; Swire, Ecker et al., 2017), it is difficult to evaluate given that it remains unpublished. There have been failures to directly replicate this study (Cameron et al., 2013), and the phenomenon has not been elicited under theoretically conducive circumstances, including a three-week delay between corrections being presented and belief being measured (Swire, Ecker et al., 2017). Furthermore, since worldview backfire effects have been demonstrated using vaccine stimuli (e.g., Nyhan et al., 2014), it is unclear whether the Skurnik et al. (2007) backfire effect was due to worldview or familiarity mechanisms. This potential misattribution also applies to Pluviano, Watt, and Della Sala (2017), Pluviano et al. (2019), and Berinsky (2017), where the backfire effects were reportedly due to familiarity mechanisms yet could have been due to worldview since the experiments exclusively used politicized information. See Appendix C for details regarding which studies found a familiarity backfire effect, which did not, and the dependent variable(s) used in each study.

There have also been recent findings that do not align with the familiarity backfire notion. For instance, simply tagging misinformation as false—with no further explanation as to why it is false—has shown to substantially reduce belief, both relative to a pre-correction within-subject baseline and in comparison to a control group who did not receive a correction at all (Ecker, O'Reilly, Reid, & Chang, 2019). Furthermore, if the familiarity backfire effect were genuine, then a practical recommendation would be to avoid repeating the misconception when presenting the correction. However, a recent meta-analysis of 10 studies found that there was no significant difference in belief updating when comparing whether or not the initial misconception was repeated within the correction (Walter & Tukachinsky, 2019).[2] Several recent studies not included in this meta-analysis

[2] For reference, we are referring to Hypothesis 3b, that the continued influence of misinformation would be stronger when corrective messages repeat the misinformation compared with those that do not repeat the misinformation. The studies included were Berinsky (2017, study 1), Ecker, Lewandowsky, Swire,

found that repeating the misconception immediately prior to the correction *facilitated* belief updating (Carnahan & Garrett, 2019; Ecker, Hogan, & Lewandowsky, 2017), and that explicitly repeating misinformation prior to the correction is more effective than only implying it (Rich & Zaragoza, 2016). Although these findings collectively oppose the familiarity backfire notion, they align well with theoretical accounts that the co-activation of the misconception and corrective information facilitates knowledge revision (Kendeou & O'Brien, 2014). It is possible that pairing the misconception and correction increases the likelihood that people notice discrepancies between the two, facilitating the integration of new information into their existing mental model (Elsey & Kindt, 2017; Kendeou, Butterfuss, Van Boekel, & O'Brien, 2017).

Finally, the illusory truth effect and familiarity backfire effect are thought to rely on the same mechanisms, and evidence suggests that the illusory truth effect can be eliminated when veracity is made salient to participants. Brashier, Eliseev, and Marsh (2020) found that when participants were simply asked to rate statements for accuracy, the illusory truth effect was wiped out. In other words, if participants knew that the item was false, the illusory truth effect was not elicited if participants were instructed to focus on accuracy both immediately and after a two-day period (also see Rapp, Hinze, Kohlhepp, & Ryskin, 2014). In sum, although repeated exposure to misinformation alone certainly increases belief, the weight of evidence suggests that this rarely, if ever, occurs when the misinformation is paired with a clear and salient correction. It remains theoretically possible that there are circumstances where the familiarity boost of the misconception outweighs the corrective element (for example, when attention is divided, Troyer & Craik, 2000), but this has not been observed empirically.

Future research can more specifically investigate how familiarity boosts that increase belief and corrections that decrease belief interact. For instance, Pennycook, Cannon, and Rand (2018) found that the increase in belief due to a single prior exposure of fake news was approximately equivalent to the reduction of belief when the fake news was accompanied by a "disputed by third-party fact-checkers" tag.

Measurement and Design Considerations

The above review suggests that backfire effects are not a robust empirical phenomenon and it could be the case that they represent an artifact of measurement error. Misinformation is still a relatively new field and more reliable measures and more powerful designs are needed to move the field ahead and determine the fate of backfire effects. Here we suggest some experimental and theoretical steps that could improve the quality of the evidence. In particular, we suggest that future studies should carefully consider measurement reliability, when possible use more powerful designs with greater internal validity, be aware of sampling and subgroup issues, and take care in linking measures with particular theories. The recommendations below could also be applicable to misinformation studies in general, rather than studies that specifically examine backfire effects.

Reliability

Reliability is defined as the consistency of a measure, that is, the degree to which a test or other measurement instrument is free of random error, yielding the same results across multiple applications to the same sample (VandenBos, 2007). Although other areas of psychology have been highly focused on measuring the reliability of their assessments (e.g., individual differences, neuropsychology, attitude research[3]), this has largely not been the case with misinformation science. A common methodological weakness in this area is the reliance on a single item to measure a belief or agreement. Single items are noisy and often have poor reliability (Jacoby, 1978; Peter, 1979), and under these conditions statistical significance may convey very little information (Loken & Gelman, 2017). Given that 81% of backfire effects found in our review of the worldview and familiarity literatures are found with single item measures, we must consider that poor item reliability could be contributing to this phenomenon. See Appendices B and C for details regarding the number of items in the measures of each study. Indeed, we found that the proportion of backfire effects observed with single item measures (37%) was significantly greater than those found in multiple item measures (8%; $Z = 2.96$, $p = .003$).

Quantifying item-level reliabilities could greatly aid in interpretation, given that a backfire effect observed on a more reliable item would have greater meaning than if found on an unreliable item. Perhaps the simplest approach to measure reliability for a single item is to include a test-retest group where participants rate their beliefs and then re-rate them after an interval has passed. This approach can be done in a control group or during a pre-baseline period in a waitlist design, if effects are expected to be extremely sample-specific. Although multi-item measures are typically more reliable than single item measures, there are occasions where single items can be sufficiently reliable (Sarstedt & Wilczynski, 2009). It is typically recommended that single-item test-retest reliability should be $\geq .70$ (Nunnally, 1978; Wanous & Hudy, 2001). Unfortunately, because so few studies in the field of misinformation have reported any measure

and Chan (2011, study 1), Nyhan and Reifler (2015), Cobb, Nyhan and Reifler, (2013, study 1), Huang (2017, study 1 and 2), Thorson (2013, study 3; 2016, study 1 and 3), and Ecker et al. (2014, study 1).

[3] Multi-item scales have long been popular in measuring attitudes (Edwards, 1983; Likert, 1974). The difference between "attitudes" and "belief" is often difficult to discern, but previous work has roughly defined attitudes as affective and beliefs as cognitive (Fishbein & Raven, 1962). We should be able to take inspiration from such attitude scales and develop items to measure how people consider the veracity of an item. For example, the belief that "listening to Mozart will make an infant more intelligent', could also be measured by asking participants whether they believe that "classical music has a unique effect on the developing prefrontal cortex". Inspiration can also be taken from studies that use "inference questions", where participants are required to use their belief in judgment tasks. For example, "If one twin listened to Mozart every night for the first 10 years of their life, and another twin was not exposed to classical music at all, how likely is it that they will have a different IQ?" or "Listening to Mozart every evening for 3 years will increase a child's IQ by what percent?" (Swire, Ecker et al., 2017).

of reliability, it is hard to know which, if any, of the items have sufficient reliability to adequately measure backfire effects.

Implementing multi-item measures could both produce more reliable measures and inspire confidence that we are measuring generalizable constructs (e.g., whether items are perceived to be important or "hot-button" issues) rather than item-specific effects. One noteworthy study by Horne, Powell, Hummel, and Holyoak (2015) incorporated a 5-item scale (reliability: $\alpha = .84$) to measure vaccine attitude changes, which correlated well with whether parents have ever refused a vaccination ($r = -0.45$). Notably, these data were subsequently reanalyzed by another group and interpreted at the individual item level because they thought the items represented separate constructs (Betsch, Korn, & Holtmann, 2015). This example not only shows that a multi-item measure can be highly reliable, but also demonstrates the challenges of creating a widely accepted multi-item measure and the current bias in the field toward analyzing individual items. In light of the replication crisis in psychology and beyond (Open Science Collaboration, 2015), all fields in the behavioral sciences have begun to focus more on measurement reliability, and a greater consideration of this issue in misinformation research could substantially improve the interpretation and replicability of our findings, particularly with regards to backfire effects.

A related issue in measuring the reliability of beliefs is that some beliefs may be stronger and more well-formulated than others. Less formulated beliefs may themselves be highly variable independent of measurement error. One approach to addressing this is to use several items to measure participants' within-belief consistency (e.g., higher consistency would indicate more formulated beliefs) as well as explicitly asking participants to rate how well-formed they perceive their beliefs to be.

A final measurement issue that could influence backfire effects is that unreliable measures have more random error and are more susceptible to regression to the mean, where extreme values at baseline testing become less extreme at follow-up testing (Bland, 2017). A regression-to-the-mean effect may be particularly problematic for individuals or subgroups in pre-post design studies who report low pre-correction belief, given that the effect could increase post-correction belief. Thus, this phenomenon could potentially result in spurious backfire effects. In Figure 1 we plot simulated data to illustrate this point. Panel A shows test-retest data for an item with poor reliability (Pearson's $r = .40$) whereas Panel B shows test-retest data for an item with good reliability (Pearson's $r = .85$). Note that at retest, data points at the extremes in the unreliable measure move more toward the mean from the line of equality (line where Time 1 = Time 2) compared to the reliable measure. Panels C and D shift these data points down 2.5 points as if a correction has been elicited. The gray area represents the "backfire zone" where post-correction belief is higher than pre-correction belief. Participants with low pre-correction belief are more likely to be found in the backfire zone when the item is unreliable (Panel C) than when it is reliable (Panel D). Though this is an oversimplified example, it shows how poor reliability and regression to the mean can give rise to spurious backfire effects in individuals and subgroups. It should be noted that effects of regression to the mean can be substantially mitigated by limiting exploratory subgroup analyses as well as including a well-matched test-retest or placebo group for comparison.

Experimental Design

In terms of design, studies of the backfire effect have varied widely, with most examining between-groups differences of correction versus no correction groups (using 5-point scales, Garrett et al., 2013; Nyhan & Reifler, 2010; 7-point scales, Wood & Porter, 2019; Zhou, 2016; percentages of participants accepting, rejecting, or unsure about the misinformation, Berinsky, 2017; or counting the mean number of references to corrected misinformation after reading a fictitious scenario, Ecker & Ang, 2019). In these studies, participants are typically randomly assigned to treatment or control, and participants' beliefs are only measured at one time point, with the experimental group being assessed after the correction. In addition to these *post-only with control* studies, a handful of studies have used within-subject pre versus post correction differences (using 11-point belief scales, Aird, Ecker, Swire, Berinsky, & Lewandowsky, 2018; Swire, Ecker et al., 2017; Swire-Thompson et al., 2019), though nearly all have lacked test-retest control groups (for an exception, see Horne et al., 2015). Other studies have used idiosyncratic approaches such as performing qualitative interviews (Prasad et al., 2009).[4]

Post-test only with control designs have an advantage in that they may be more practically feasible, often only requiring a single testing session. Another advantage of this design is that researchers are able to test belief without a further familiarity boost, which is potentially important for studies attempting to examine the familiarity backfire effect. Post-test only with control designs are also thought to limit carryover effects associated with pre-post designs, although it is questionable whether carryover effects are a concern in misinformation studies. If carryover effects were problematic, participants in pre-post studies would provide post-correction responses that are similar to their initial response, and the belief change observed in these designs would be significantly smaller than post-test only with control designs. However, effect sizes of belief change in pre-post studies are similar in magnitude to post-test only with control designs, suggesting that the impact of carryover effects is likely minimal. In fact, Ecker et al. (2019) found larger decreases in belief in false claims if the manipulation was within-subjects rather than between subjects. Furthermore, effect sizes for belief change in pre versus post-correction studies are typically large, *especially* in conditions where there is no delay between pre and posttest, where one would expect carryover effects to be most pronounced

[4] Prasad developed a technique called "challenge interviews" where interviewers presented participants with substantive challenges to their political opinions. They themselves do not claim their findings were a backfire effect in their paper but are frequently cited in support of the worldview backfire phenomenon. They found that the most popular strategy was "attitude bolstering", bringing facts that support one's position to mind without directly refuting the contradictory information. Belief change was not measured.

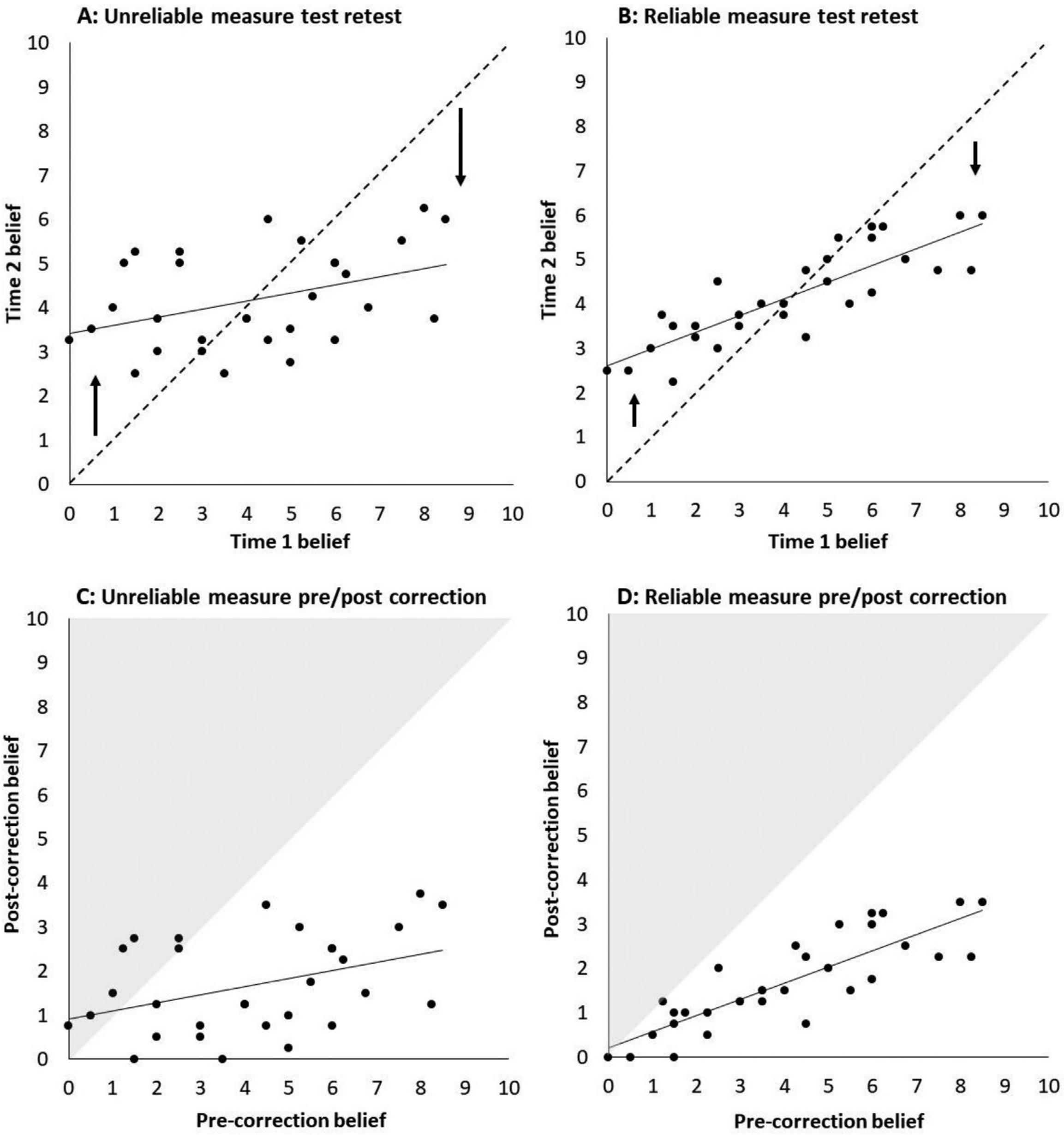

Figure 1. Simulated data of 30 participants on a misinformation item. Panel A illustrates test-retest data for an item with poor reliability ($r = .40$) and Panel B for an item with acceptable reliability ($r = .85$). The dotted lines in Panel A and B represent the line of equality, on which all data points would lie if the measurements at Time 1 and Time 2 were identical. Note greater regression-to-the-mean effects in the unreliable data than the reliable data, indicated by the arrows. Panels C and D shift these data points down 2.5 points as if a correction has been elicited. The gray area represents the "backfire zone." Panel C shows pre/post data demonstrating that subjects reporting low pre-correction beliefs may be more likely to result in spurious backfire effects if the measures are unreliable.

(e.g., $\eta_p^2 = .53$, Swire-Thompson et al., 2019; $\eta_p^2 = .62$, Aird et al., 2018).

An important issue with designs such as the post-test only with control, is that even with randomization the groups may not be adequately matched at baseline (Morris, 2008). This issue may be particularly problematic with smaller sample sizes. The prevalence of this problem is hard to assess because many studies fail to report important demographic and political characteristics and rarely compare these variables between experimental and control groups. It is easy to imagine small group differences in demographics and political polarization producing the appearance of backfire effects. Pre-post designs are a viable alternative to post-test only designs and are not as affected by sampling issues. However, pre-post designs without a control group (e.g., Swire, Ecker et al., 2017) could also potentially suffer from problems related to repeated testing effects such as regression to the mean (Vickers & Altman, 2001), which could drive backfire effects, particularly at the subgroup level (see Figure 1).

A more powerful design that can overcome many of these issues is a *pre-post control group design,* where participants are randomly assigned to intervention or control conditions, and participants are measured both before and after the intervention or control is administered (Schulz, Altman, & Moher, 2010). This design is common in clinical intervention studies, and compared to post-test only between-subject designs, it offers a boost in statistical power. Further, because the experimental manipulation is within-subjects, the internal validity of this design does not solely depend on random assignment (Charness, Gneezy, & Kuhn, 2012). One important question for this design with regards to the backfire effect is what the best control condition would be. Though a test-retest control is a starting point, a more sophisticated approach would be to employ one or multiple placebo conditions (e.g., where participants are given information related to the misinformation that does not correct nor confirm it). The only study that we are aware of that has used this design in the context of backfire effects is Horne et al. (2015). They compared vaccine attitudes in 315 adults before and after random assignment to either (a) autism correction, (b) disease risk information, or (c) reading an unrelated scientific vignette (control condition). Notably, they did not find backfire effects at the group level. When exploring subgroups, those with the least favorable attitudes toward vaccines were the *most* receptive to change. However, those with the most favorable attitudes to vaccines did show a backfire, which the authors interpreted as regression to the mean. Though this type of design may be more participant- and resource-demanding than post-only with control or pre-post designs, it could help provide a more powerful evaluation of the possible presence of backfire effects.

We finally turn to demand characteristics and whether participants' expectations for how they are meant to behave facilitate backfire effects (Orne, 1959). Demand characteristics generally lead participants to be a "good subject," encouraging them to behave in a manner that confirms the hypothesis of the experimenter (Nichols & Maner, 2008). If the participant does receive cues that the experiment is about the efficacy of belief updating, they are likely to further *reduce* their belief after viewing a correction, rather than report increasing their belief. The only study, to our knowledge, that explicitly asked subjects about the purpose of the experiment in a debriefing questionnaire was conducted by Ecker, Lewandowsky, and Tang (2010), and they found that virtually all participants did indeed correctly assume that the experiment was about memory updating. It is nonetheless important that future studies quantify the extent to which demand characteristics influence misinformation experiments in general. Should future investigations deem them problematic, one method to reduce demand characteristics is to blind participants to the goals of the study and, for in-lab studies, blind experimenters (see Orne, 2009).

Sampling and Subgroup Issues

Another step forward for backfire studies is to be more aware of sampling and subgroup issues because the subgroups in which backfire effects have been found vary substantially. As we previously noted, the internal validity of between-groups post-test only designs can be seriously undercut by demographic differences between the groups, and more thorough between-groups demographic comparisons in these studies is essential. Further, though some studies that test for backfire effects have used previously defined subgroups (e.g., Ecker & Ang, 2019; Haglin, 2017; Wood & Porter, 2019), some of the subgroup analyses reported may have been post hoc. Post hoc subgroup analyses have been harshly criticized in the clinical literature (Wang, Lagakos, Ware, Hunter, & Drazen, 2007) because it is often unclear how many are performed and whether they are motivated by inspection of the data. Thus, in future studies, subgroup analyses derived from data inspection should be explicitly identified as exploratory (as done by Nyhan & Reifler, 2010), and subgroup analyses should be pre-specified, or better yet, pre-registered.

Practical Recommendations

Regarding the worldview backfire effect, fact-checkers can rest assured that it is *extremely unlikely* that, at the broader group level, their fact-checks will lead to increased belief in the misinformation. Meta-analyses have clearly shown that corrections are generally effective and backfire effects are not the norm (e.g., Chan, Jones, Hall Jamieson, & Albarracín, 2017; Walter & Murphy, 2018). Furthermore, given that research has yet to systematically show backfire effects in the same subgroups, practitioners should not avoid giving corrections to any specific subgroups of people. Fact-checkers can therefore focus on other known issues such as getting the fact-checks to the individuals who are most likely to be misinformed.

Regarding the familiarity backfire effect, avoiding the repetition of the original misconception within the correction appears to be unnecessary and could even hinder corrective efforts (Ecker et al., 2017; Kendeou & O'Brien, 2014). We therefore instead suggest designing the correction first and foremost with clarity and ease of interpretation in mind. Although the familiarity backfire effect lacks evidence, we must be aware that the illusory truth effect *in the absence* of corrections or veracity judgments is extremely robust. Therefore, when designing a correction, the misinformation should always be *clearly and saliently* paired with the corrective element, and needless repetitions of the misconception should still be avoided. For instance, given that many individuals do not read further than headlines (Gabielkov, Ramachandran, Chaintreau, & Legout, 2016), the misconception should not be described in the headline alone with the correction in smaller print in the text below (Ecker, Lewandowsky, Chang, & Pillai, 2014; Ecker, Lewandowsky, Fenton, & Martin, 2014). Adding the corrective element within the headline itself, even if it is simply a salient "myth" tag associated with the misconception, can be considered good practice.

Future Research

Although improvements in both experimental measures and designs are important, Oberauer and Lewandowsky (2019) highlight that another cause of poor replicability is weak logical links between theories and empirical tests. Future research could more explicitly manipulate key factors presumed to influence belief updating, whether it be fluency, perceived item importance, strength of belief, complexity of the item wording, order of corrective elements, internal counter-arguing, source of the message, or participants' communicating disagreement with the correction. Focusing on theoretically meaningful factors could help to better isolate the potential mechanisms behind backfire effects or the continued influence effect in general. Furthermore, it would be beneficial to be aware of other competing factors to avoid confounds. For example, when investigating the effects of familiarity, one could avoid exclusively using issues presumed to elicit worldview backfire effects (e.g., vaccines, Skurnik et al., 2007). Additionally, given that responses to corrections are likely heterogeneous, it would be beneficial to use a wide variety of issues in experiments that vary on theoretically meaningful criteria to dissociate when backfire effects occur and when they do not.

Future research should also empirically investigate common recommendations that stem from the familiarity backfire effect notion which have yet to be thoroughly examined. For example, it is unclear whether belief updating is fostered by presenting a "truth sandwich" to participants, stating the truth twice with the falsehood between (Sullivan, 2018). Preliminary findings suggest that a "bottom-loaded" correction, which first states the misconception followed by two factual statements, could be more effective than the truth sandwich (Anderson, Horton, & Rapp, 2019), although further research is required prior to firm recommendations being made.

Finally, there are additional occasions where corrections could be counter-productive that require empirical investigation. For instance, correcting facts in public political debate might not always be advisable, because it involves the acceptance of someone else's framing, allowing the person who promulgated the original falsehood to set the agenda (Lakoff, 2010; Lewandowsky, Ecker, & Cook, 2017). Furthermore, broadcasting a correction where few people believe in the misconception could be a legitimate concern, since the correction may spread the misinformation to new audiences (Kwan, 2019; Schwarz et al., 2016). For example, if the *BBC* widely publicized a correction to a misconception that its readership never believed to begin with, it will not reap the benefits of belief reduction, and those who do not trust this source may question its conclusion. The next crucial step is to examine such instances with real-world scenarios on social media or fact-checking websites.

Conclusion

In today's fast-paced information society it is extremely important to understand the efficacy of corrections, the exact circumstances under which they have no impact, or even backfire. The current state of the literature calls into question the notion of the backfire effect and more rigorous studies are needed to determine if there are circumstances when these effects reliably occur. Indeed, given the current coronavirus pandemic and the rampant misinformation that has accompanied it, understanding the parameters of misinformation correction is particularly crucial. In sum, the current review suggests that backfire effects are not a robust empirical phenomenon, and more reliable measures, powerful designs, and stronger links between experimental design and theory could greatly help move the field ahead.

Author Contributions

BST conceived of the idea and wrote the majority of the manuscript. JD contributed sections to the manuscript, particularly the measurement and design considerations. BST, JD, and DL edited the manuscript.

Conflict of Interest Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This project was funded by a Hewlett Packard Foundation grant. All authors acknowledge that the material presented in this manuscript has not been previously published and is not under consideration by any other journal. We thank Nicholas Miklaucic for proofreading and verifying the information in Appendices A, B, and C.

Appendix A. Papers frequently cited in support of backfire effects that do not meet the criteria of a backfire effect

Article	Type of backfire	Reason for exclusion
Skurnik et al. (2005)	Familiarity	This study did not meet the criteria of a backfire effect since post-correction belief is not compared to a pre-correction or no-correction baseline. They considered a backfire to be misremembering more false items as true than true items to be false. For a critique, see Swire, Ecker et al. (2017).
Weaver, Garcia, Schwarz, & Miller (2007)	Familiarity	This study is regarding the illusory truth effect, but since it does not provide corrections to participants it cannot comment on the backfire effect.
Prasad et al. (2009)	Worldview	This study consisted of qualitative interviews and belief change was not measured.
Peter and Koch (2016)	Familiarity	This study did not meet the criteria of a backfire effect since post-correction belief is not compared to a pre-correction or no-correction baseline. They considered the backfire to be misremembering more false items as true than true items as false.
Holman and Lay (2019)	Worldview	We excluded this study because even though a backfire effect is reported amongst Republicans when provided with a non-partisan fact check, these findings are not significant at the $p < .05$ level.

Appendix B. Worldview backfire effect literature

Article	Topic	Sample	Dependent variable(s)	Number of items in DV	Did a backfire occur? / In what subgroup
Nyhan and Reifler (2010)	Exp 1: Weapons of mass destruction	130 university students	Agreement with claim	1 item	Yes: Conservatives
	Exp 2: Weapons of mass destruction, tax cuts, and stem cell research.	195 university students	Agreement with weapons of mass destruction claim	1 item	No
			Agreement with tax cuts claim	1 item	Yes: Conservatives
			Agreement with stem cell research claim	1 item	No
Hart and Nisbet (2012)	Climate change	240 community sample	Support for government action	3 items (composite)	Yes: Republicans
Garrett et al. (2013)	Islamic cultural center	750 representative online participants	Belief in claim	1 item	No
Nyhan et al. (2013)	Palin death panel	945 online participants	Belief in claim	1 item	Yes: Politically knowledgeable Palin supporters
		936 online participants	Approval of Affordable Care Act	1 item	Yes: Politically knowledgeable Palin supporters
Ecker, Lewandowsky, Fenton, & Martin (2014)	Exp1: Fictitious scenario about a store robbery	144 university students	Inference questions (references to the misinformation)	10 items (composite)	No
	Exp 2: Fictitious scenario about an Aboriginal hero	100 university students	Inference questions (references to the misinformation)	10 items (composite)	No
Nyhan et al. (2014)	Vaccines	1736 parents	Agreement with claim	1 item	No
		1746 parents	MMR side effect	1 item	No
		1751 parents	Vaccine intent	1 item	Yes: Individuals with least favorable vaccine attitudes
Weeks and Garrett (2014)	Political candidates	600 community sample	Belief in claims	10 items: (2 composites of 4 false items)	No
Horne et al. (2015)	Vaccines	315 online participants (137 parents)	Attitudes toward vaccines	5 items (composite)	No
Nyhan and Reifler (2015)	Vaccines	995 representative online participants	Accuracy of claim	1 item	No
		997 representative online participants	General belief about safety of flu vaccines	1 item	No
		998 representative online participants	Intent to vaccinate	1 item	Yes: Individuals with high levels of concern about vaccine side effects
Weeks (2015)	Political misinformation	768 online participants	Belief	8 items (2 composites of 4 items)	No
Zhou (2016)	Climate change	475 online participants	Attitudes toward governmental action against climate change	3 items (composite)	Yes: Republicans with high political interest
			Likelihood to take personal action	1 item	Yes: Republicans with high political interest
			Attitudinal ambivalence	3 items (composite)	No
Trevors, Muis, Pekrun, Sinatra, & Winne (2016)[a]	Genetically modified foods	120 university students	Knowledge	10 items (composite)	No

Haglin (2017)	Vaccines	474 online participants	Accuracy of claim	1 item	No
			General belief about safety of flu vaccines	1 item	No
			Intent to vaccinate	1 item	No
Swire, Berinsky et al. (2017)	Exp 1: Political misinformation	1776 online participants	Belief	8 items (2 composites of 4 items)	No
	Exp 2: Political misinformation	960 representative online participants	Belief	6 items (2 composites of 3 items)	No
Aird et al. (2018)	Political misinformation	370 university students and online participants	Accuracy of claim	8 items (2 composites of 4 items)	No
Ecker and Ang (2019)	Political misinformation	162 university students	Inference questions (references to the misinformation)	10 items (composite)	Yes: Right-wing participants; attentive sample only
Nyhan et al. (2019)	Exp 1: Rising crime	3420 online participants	Belief	1 item	No
	Exp 2: Loss of manufacturing jobs	825 online participants	Belief	1 item	No
Schmid and Betsch (2019)	Exp 1: Vaccines	112 university students	Intent to vaccinate	1 item	No
			Attitudes toward vaccination	3 items (composite)	No
	Exp 2: Vaccines	158 representative online participants	Intent to vaccinate	1 item	No
			Attitudes toward vaccination	3 items (composite)	No
	Exp 3: Vaccines	198 representative online participants	Intent to vaccinate	1 item	No
			Attitudes toward vaccination	3 items (composite)	No
	Exp 4: Vaccines	227 online participants	Intent to vaccinate	1 item	No
			Attitudes toward vaccination	3 items (composite)	No
	Exp 5: Climate	148 recruited via mailing list and social media advertising	Intention to act against climate change	7 items (composite)	No
			Attitudes toward climate	3 items (composite)	No
	Exp 6: Vaccines	921 online participants	Intent to vaccinate	1 item	No
			Attitudes toward vaccination	3 items (composite)	No
Swire-Thompson et al. (2019)	Political misinformation	1500 online participants	Belief	8 items (composites of 4 true and 4 false)	No
Wood and Porter (2019)	Exp1: Political misinformation	3127 online participants	Agreement with claim	8 items (analyzed individually)	No
	Exp 2: Political misinformation	2801 online participants	Agreement with claim	16 items (analyzed individually)	No
	Exp 3: Political misinformation	977 online participants	Agreement with correction	7 items (analyzed individually)	No
	Exp 4: Political misinformation	1333 online participants	Agreement with correction	18 items (analyzed individually *and* 3 composites made of 6 items each)	No
	Exp 5: Political misinformation	2019 online participants	Agreement with correction	6 statements (analyzed individually *and*in 3 composites)	No

Note. Dependent variables and studies that report backfire effects are highlighted in gray. We also highly recommend viewing Guess and Coppock (2018), although the findings are omitted from this table because they presented counter-attitudinal information rather than corrective information to participants.

[a] Trevors et al. (2016) found that self-concept negatively predicted attitudes after reading refutation text more than expository text, which the authors concluded as evidence of a backfire effect. We exclude the attitude dependent variable from this table because there was no between-subject control group and the pre-post attitude tests could not be compared.

Appendix C. Familiarity backfire effect literature

Article	Topic	Sample	Dependent variable(s)	Number of items in DV	Did a backfire occur? / In what subgroup
Skurnik et al. (unpublished manuscript, 2007)[a]	Vaccines	Unknown	Intent to vaccinate	Appears to be 1 item (although exact details unavailable).	Yes: 30-minute delay after corrections presented
Cameron et al. (2013)	Vaccines	105 community sample	Knowledge	15 items (composite)	No
			Accuracy of veracity recall	8 items (composite)	No
Rich and Zaragoza (2016)	Exp 1: Fictitious story about a warehouse fire	328 universitystudents	Inference questions (references to the misinformation)	10 items (composite)	No
	Exp 2: Fictitious story about a warehouse fire	338 university students and online participants	Inference questions (references to the misinformation)	10 items (composite)	No
Berinsky (2017)[b]	Exp 1: Death panels	699 representative online participants	Belief	2 items (analyzed individually)	No
			Support for health care plan	1 item	No (Backlash mentioned, but Democratic correction is not statistically different from control)
Ecker et al. (2017)	Fictional news reports (e.g. about a wildfire)	60 university students	Inference questions (references to the misinformation)	7 items (composite)	No
Pluviano et al. (2017)	Vaccines	120 university students	Agreement that vaccines cause autism	1 item	Yes: Myth vs Fact correction condition, after 7 days
			Belief in vaccine side effects	1 item	Yes: Fear correction, immediately and after 7 days
			Intent to vaccinate	1 item	Yes: Myth vs Fact correction condition, after 7 days
Swire, Ecker et al. (2017)	Exp 1: Range of topics	93 university students	Belief	40 items (composite of 20 myths and 20 facts)	No
	Exp 2: Wide range of topics	109 older adult community sample	Belief	40 items (composite of 20 myths and 20 facts)	No
Carnahan and Garrett (2019)	Exp 1: Electronic health records	344 online participants	Belief	7 items (composite)	No
	Exp 2: Genetically modified foods	486 online participants	Belief	7 items (composite)	No
Ecker et al. (2019)	Exp 1: Range of topics	531 online participants	Belief	12 items (composite of 6 false and 6 true)	No
	Exp 2: Range of topics	369 online participants	Belief	12 items (composite of 6 false and 6 true)	No
Pennycook et al. (2018)[c]	Exp 2: Various fake news headlines	949 online participants	Accuracy	24 items (4 composites of 6 items)	No
	Exp 3: Various fake news headlines	940 online participants	Accuracy	24 items (6 composites of 4)	No
Pluviano et al. (2019)	Vaccines	60 parents	Agreement that vaccines cause autism	1 item	Yes: After 7 days
			Belief in vaccine side effects	1 item	Yes: After 7 days
			Intent to vaccinate	1 item	No

Note. Dependent variables and studies that report backfire effects are highlighted in gray.

[a] In addition to intent to vaccinate, Skurnik et al. (2007) also considered the misremembering of more myths thought to be true than facts thought to be false to stem from famili arity mechanisms. We have excluded this element from the table because they do not meet the criteria of a backfire effect since it is not in comparison to a pre-correction or no-correction baseline.

[b] Berinsky (2017)'s second experiment found the trend th at repetition of the misinformation prior to the correction made respondents less likely to reject it than if the misinformation was not repeated. However, we exclude this study since there is no pre-correction or no-correction baseline.

[c] We exclude the first experiment from Pennycook et al. (2018) it was investigating the illusory truth effect. Since it did not provide corrections to participants it cannot comment on the backfire effect.

References

Aird, M. J., Ecker, U. K., Swire, B., Berinsky, A. J., & Lewandowsky, S. (2018). Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *Royal Society Open Science*, *5*, 180593.

Anderson, E. R., Horton, W., & Rapp, D. (2019). Hungry for the truth: Evaluating the utility of "truth sandwiches" as refutations. [Conference presentation]. *Annual meeting of the society for text & discourse*.

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*, 446–458. http://dx.doi.org/10.1037/0096-3445.121.4.446

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, *47*, 241–262.

Betsch, C., Korn, L., & Holtmann, C. (2015). Don't try to convert the antivaccinators, instead target the fence-sitters. *Proceedings of the National Academy of Sciences*, *112*, E6725–E6726.

Bland, M. (2017). Errors of measurement: Regression toward the mean. In M. Allen (Ed.), *The SAGE encyclopedia of communication research methods*. SAGE Publications.

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

Cameron, K. A., Roloff, M. E., Friesema, E. M., Brown, T., Jovanovic, B. D., Hauber, S., et al. (2013). Patient knowledge and recall of health information following exposure to "facts and myths" message format variations. *Patient Education and Counseling*, *92*, 381–387.

Carnahan, D., & Garrett, R. K. (2019). Processing style and responsiveness to corrective information. *International Journal of Public Opinion Research*.

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*(11), 1531–1546.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*, 1–8.

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*(3), 307–326.

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.

Cook, J., & Lewandowsky, S. (2012). *The debunking handbook*.. http://www.skepticalscience.com/docs/Debunking_Handbook.pdf

Ecker, U. K. H., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology*, *40*, 241–260. http://dx.doi.org/10.1111/pops.12494

Ecker, U. K., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory and Cognition*, *42*, 292–304.

Ecker, U. K., 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.

Ecker, U. K., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, *20*, 323.

Ecker, U. K., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition*, *42*, 292–304.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570–578.

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

Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2019). The effectiveness of short-format refutational fact-checks. *British Journal of Psychology*.

Edwards, A. L. (1983). *Techniques of attitude scale construction*. Irvington Publishers Inc.

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

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*.

Fishbein, M., & Raven, B. H. (1962). The AB scales: An operational definition of belief and attitude. *Human Relations*, *15*, 35–44.

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

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter? *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*.

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: Undermining corrective effects. *Journal of Communication*, *63*, 617–637.

Gemberling, T. M., & Cramer, R. J. (2014). Expert testimony on sensitive myth-ridden topics: Ethics and recommendations for psychological professionals. *Professional Psychology: Research and Practice*, *45*, 120.

Guess, A., & Coppock, A. (2018). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 1–19. http://dx.doi.org/10.1017/S0007123418000327

Haglin, K. (2017). The limitations of the backfire effect. *Research & Politics*, http://dx.doi.org/10.1177/2053168017716547

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication how motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*, 701–723.

Holman, M. R., & Lay, J. C. (2019). They see dead people (voting): Correcting misperceptions about voter fraud in the 2016 US presidential election. *Journal of Political Marketing*, *18*, 31–68.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences*, *112*, 10321–10324.

Huang, H. (2017). A war of (mis) information: The political effects of rumors and rumor rebuttals in an authoritarian country. *British Journal of Political Science*, *47*(2), 283–311.

Jacoby, J. (1978). Consumer research: How valid and useful are all our consumer behavior research findings? A state of the art review1. *Journal of Marketing*, *42*, 87–96.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When discredited information in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 1420–1436.

Kendeou, P., Butterfuss, R., Van Boekel, M., & O'Brien, E. J. (2017). Integrating relational reasoning and knowledge revision during reading. *Educational Psychology Review*, *29*, 27–29.

Kendeou, P., & O'Brien, E. J. (2014). The Knowledge Revision Components (KReC) framework: Processes and mechanisms. In D. Rapp, & J. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. Cambridge: MIT.

Kowalski, P., & Taylor, A. K. (2017). Reducing students' misconceptions with refutational teaching: For long-term retention, comprehension matters. *Scholarship of Teaching and Learning in Psychology*, *3*, 90.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Kwan, V. (2019). Responsible reporting in an age of information disorder. *First Draft Report*, https://firstdraftnews.org/wpcontent/uploads/2019/10/Responsible_Reporting_Digital_AW-1.pdf?x20994.

Lakoff, G. (2010). *Moral politics: How liberals and conservatives think*. University of Chicago Press.

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. http://dx.doi.org/10.1177/1529100612451018

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

Likert, R. (1974). A method of constructing an attitude scale. *Scaling: A Sourcebook for Behavioral Scientists*, 233–243.

Lilienfeld, S. O., Marshall, J., Todd, J. T., & Shane, H. C. (2014). The persistence of fad interventions in the face of negative scientific evidence: Facilitated communication for autism as a case example. *Evidence-Based Communication Assessment and Intervention*, *8*, 62–101.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, *135*, 151–166.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 1–22.

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

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: A randomized trial. *Pediatrics*, *133*, e835–e842. http://dx.doi.org/10.1542/peds.2013-2365

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*, 459–464. http://dx.doi.org/10.1016/j.vaccine.2014.11.017

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, *51*, 127–132.

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596–1618.

Orne, M. T. (1959). The nature of hypnosis: Artifact and essence. *The Journal of Abnormal and Social Psychology*, *58*, 277.

Orne, M. T. (2009). Demand characteristics and the concept of quasi-controls. In R. Rosenthal, & R. L. Rosnow (Eds.), *Artifacts in Behavioral Research* (pp. 110–137).

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*.

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, *16*, 6–17.

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.

Prasad, M., Perrin, A. J., Bezila, K., Hoffman, S. G., Kindleberger, K., Manturuk, K., et al. (2009). "There must be a reason": Osama, Saddam, and inferred justification. *Sociological Inquiry*, *79*, 142–162.

Pluviano, S., Watt, C., & Della Sala, S. (2017). Misinformation lingers in memory: failure of three pro-vaccination strategies. *PLoS One*, *12*(7), Article e0181640.

Pluviano, S., Watt, C., Ragazzini, G., & Della Sala, S. (2019). Parents' beliefs in misinformation about vaccines are strengthened by pro-vaccine campaigns. *Cognitive Processing*, 1–7.

Rapp, D. N., Hinze, S. R., Kohlhepp, K., & Ryskin, R. A. (2014). Reducing reliance on inaccurate information. *Memory & Cognition*, *42*, 11–26.

Rich, P. R., & Zaragoza, M. S. (2016). The continued influence of implied and explicitly stated misinformation in news reports. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 62.

Sarstedt, M., & Wilczynski, P. (2009). More for less? A comparison of single-item and multi-item measures. *Die Betriebswirtschaft*, *69*, 211.

Schmid, P., & Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, *1* http://dx.doi.org/10.1038/s41562-019-0632-4

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*, 18.

Schwarz, N., Sanna, L. J., Skurnik, I., & Yoon, C. (2007). *Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns*. pp. 127–161. *Advances in Experimental Social Psychology* (Vol. 39) , http://linkinghub.elsevier.com/retrieve/pii/S006526010639003X.

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.

Science, A. B. C. (2017). *What is the backfire effect*. Available from: https://www.scienceabc.com/humans/what-is-the-backfire-effect-confirmation-bias-psychology.html.

Skurnik, I., Yoon, C., & Schwarz, N. (2007). Education about flu can reduce intentions to get a vaccination. Unpublished manuscript.

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

Swire-Thompson, B., Ecker, U. K. H., Lewandowsky, S., & Berinsky, A. J. (2019). They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology*.

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*(4), 713–724.

Sullivan, M. (2018). *Instead of Trump's propaganda, how about a nice 'truth sandwich'?* The Washington Post. https://www.washingtonpost.com/lifestyle/style/instead-of-trumps-propaganda-how-about-a-nice-truth-sandwich/2018/06/15/80df8c36-70af-11e8-bf86-a2351b5ece99_story.html

Taber, C. S., & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science*, *50*, 755–769. http://dx.doi.org/10.1111/j.1540-5907.2006.00214.x

Thorson, E. (2013). Belief echoes: The persistent effects of correction misinformation. In *Unpublished dissertation*. University of Pennsylvania.

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

Trevors, G. J., Muis, K. R., Pekrun, R., Sinatra, G. M., & Winne, P. H. (2016). Identity and epistemic emotions during knowledge revision: A potential account for the backfire effect. *Discourse Processes*, *53*, 339–370.

Troyer, A. K., & Craik, F. I. M. (2000). The effect of divided attention on memory for items and their context. *Canadian Journal of Experimental Psychology*, *54*, 161–170.

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

VandenBos, G. R. (Ed.). (2007). *APA dictionary of psychology*. American Psychological Association.

Vickers, A. J., & Altman, D. G. (2001). Analysing controlled trials with baseline and follow up measurements. *BMJ*, *323*, 1123–1124.

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.

Walter, N., & Tukachinsky, R. (2019). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., & Drazen, J. M. (2007). Statistics in medicine—Reporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, *357*, 2189–2194.

Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods*, *4*, 361–375.

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.

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

Weeks, B. E., & Garrett, R. K. (2014). Electoral consequences of political rumors: Motivated reasoning, candidate rumors, and vote choice during the 2008 U.S. Presidential Election. *International Journal of Public Opinion Research*, *26*, 401–422. http://dx.doi.org/10.1093/ijpor/edu005

Wells, C., Reedy, J., Gastil, J., & Lee, C. (2009). Information distortion and voting choices: The origins and effects of factual beliefs in initiative elections. *Political Psychology*, *30*, 953–969.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*, 135–163.

Zhou, J. (2016). Boomerangs versus javelins: How polarization constrains communication on climate change. *Environmental Politics*, *25*, 788–811.

Received 14 January 2020;
received in revised form 20 June 2020;
accepted 21 June 2020
Available online 2 September 2020