The Backfire Effect After Correcting Misinformation Is Strongly Associated With Reliability

Briony Swire-Thompson, Nicholas Miklaucic, John P. Wihbey, David Lazer, and Joseph DeGutis

The Backfire Effect After Correcting Misinformation Is Strongly Associated With Reliability

Briony Swire-Thompson[1, 2], Nicholas Miklaucic[1], John P. Wihbey[3], David Lazer[1, 2], and Joseph DeGutis[4, 5]
[1] Network Science Institute, Northeastern University
[2] Institute of Qualitative Social Science, Harvard University
[3] Journalism and Media Innovation, Northeastern University
[4] Boston Attention and Learning Laboratory, VA Boston Healthcare System, Boston, Massachussets, United States
[5] Department of Psychiatry, Harvard Medical School

The backfire effect is when a correction increases belief in the very misconception it is attempting to correct, and it is often used as a reason not to correct misinformation. The current study aimed to test whether correcting misinformation increases belief more than a no-correction control. Furthermore, we aimed to examine whether item-level differences in backfire rates were associated with test–retest reliability or theoretically meaningful factors. These factors included worldview-related attributes, including perceived importance and strength of precorrection belief, and familiarity-related attributes, including perceived novelty and the illusory truth effect. In 2 nearly identical experiments, we conducted a longitudinal pre/post design with $N = 388$ and 532 participants. Participants rated 21 misinformation items and were assigned to a correction condition or test-retest control. We found that no items backfired more in the correction condition compared to test-retest control or initial belief ratings. Item backfire rates were strongly negatively correlated with item reliability ($\rho = -.61/-.73$) and did not correlate with worldview-related attributes. Familiarity-related attributes were significantly correlated with backfire rate, though they did not consistently account for unique variance beyond reliability. While there have been previous papers highlighting the nonreplicable nature of backfire effects, the current findings provide a potential mechanism for this poor replicability. It is crucial for future research into backfire effects to use reliable measures, report the reliability of their measures, and take reliability into account in analyses. Furthermore, fact-checkers and communicators should not avoid giving corrective information due to backfire concerns.

Keywords: misinformation, reliability, belief updating, the backfire effect

Supplemental materials: https://doi.org/10.1037/xge0001131.supp

In today's information-based society, people are continuously presented with misinformation that needs to be corrected (Lewandowsky et al., 2012). It is extremely important that we know if corrections are effective and whether there are circumstances under which they are likely to be ineffective, or even make matters worse. The backfire effect is when belief in a misconception is greater after a correction has been presented in comparison to a precorrection or no-correction baseline (see Swire-Thompson, DeGutis, & Lazer, 2020 for a review). Although there have been multiple failures to replicate (for instance, Wood & Porter, 2019), the backfire effect is still hotly debated (Jerit & Zhao, 2020), with some researchers maintaining that it is a genuine phenomenon (e.g., Autry & Duarte, 2021; Schwarz & Jalbert, 2020). The current consensus is that the backfire effect only occurs in limited

circumstances (Lewandowsky et al., 2020; Nyhan, 2021). The present study sought to more thoroughly characterize the backfire effect by examining if it occurs more than a no-correction control condition and if item-level variation in backfire rate is accounted for by either (a) an item's test–retest reliability or (b) theoretically meaningful factors proposed to underlie backfire effects.

There are two types of backfire effects that have become popular in the literature: the worldview backfire effect and the familiarity backfire effect. They both result in increased belief postcorrection, but they are presumed to occur due to different psychological mechanisms. The worldview backfire effect is said to be elicited when a correction challenges an individual's belief system (Cook & Lewandowsky, 2012) and is commonly suggested to be more likely to occur when the information is *important* to the individual or when the person believes in the information strongly (Flynn et al., 2017; Lewandowsky et al., 2012; Nyhan & Reifler, 2010; Nyhan & Reifler, 2015). Although there was initially promising evidence for the worldview backfire effect (for instance, Nyhan & Reifler, 2010), there have been failures to find or replicate this effect (e.g., Guess & Coppock, 2018; Ecker et al., 2021; Haglin, 2017; Nyhan et al., 2020; Schmid & Betsch, 2019; Swire, Berinsky, et al., 2017; Weeks & Garrett, 2014; Wood & Porter, 2019). Furthermore, to our knowledge worldview backfire effects have exclusively been found in subgroup analyses, and the type of subgroups have been inconsistent (Swire-Thompson, DeGutis, & Lazer, 2020). In addition to these issues, previous work has yet to explicitly investigate whether the posited worldview mechanisms—the perceived importance of the information or the strength of one's belief prior to correction—are associated with a greater likelihood of backfire effects.

In contrast to the mechanisms of the worldview backfire effect, a familiarity backfire effect is said to occur when misinformation is repeated within the retraction (Schwarz et al., 2007). For example, the statement "The COVID-19 vaccine will NOT alter your DNA" seeks to dispel the misinformation, yet the repetition of "COVID-19 vaccine" and "altered DNA", is proposed to make this inaccurate link more familiar. The familiarity backfire effect is thought to result from people being more likely to believe that repeated information is true, a phenomenon called the illusory truth effect (Begg et al., 1992; Fazio et al., 2019). The illusory truth effect and the familiarity backfire effect are thought to be driven by the same psychological mechanism of processing fluency (see Unkelbach et al., 2019 for a review), even though the former occurs in the *absence* of corrective information and the latter in the *presence* of corrective information. There are many factors that influence processing fluency, from the use of easy- or hard-to-pronounce words (Alter & Oppenheimer, 2006) to the degree to which the statement rhymes (McGlone & Tofighbakhsh, 2000). If processing fluency underlies both the illusory truth effect and the familiarity backfire effect, this would predict that items prone to one effect would also be prone to the other. However, to date, this has never been tested.

The most commonly cited evidence for the familiarity backfire effect is an unpublished article (Skurnik et al., 2007 as cited by Schwarz et al., 2007). The authors found that participants who read a "myths versus facts" pamphlet debunking vaccine misconceptions had less favorable attitudes toward vaccines compared to a control group who never saw the pamphlet. While subsequent research has found little evidence for the phenomenon (e.g., Cameron et al., 2013; Ecker et al., 2017; Swire, Ecker, et al., 2017), one potentially moderating factor is that novel misinformation could be theoretically more susceptible to backfire effects than familiar misinformation (as suggested by Ecker,

O'Reilly, et al., 2020). Compared to well-known items, a correction applied to novel items will disproportionately boost the familiarity of the misconception from a no-exposure baseline. Ecker, Lewandowsky, et al. (2020) tested this hypothesis at the group level. Although a small backfire effect was observed in the first experiment, they found that corrections of novel misinformation generally did not lead to stronger belief in the misinformation than a control group who were never exposed to the false claims or corrections. However, the question still remains whether the novelty of information makes it *more likely* for backfire effects to occur.

Before examining worldview- and familiarity-related characteristics of items, it is important to quantify whether item differences in backfire rate vary with regards to reliability. Interest in measurement reliability, the consistency of a measure and the degree to which it is free of random error (VandenBos, 2007), has steadily grown in psychology and neuroscience research in recent years (e.g., Zuo et al., 2019), likely due to the replication crisis in these fields. At best, unreliable measures add noise and complicate the interpretation of effects observed. At worst, unreliable measures can produce statistically significant findings that are spurious artifacts (Loken & Gelman, 2017). A significant drawback of prior misinformation research is that experiments investigating backfire effects have typically not reported the reliability of their measures (for an exception, see Horne et al., 2015). Due to random variation or regression to the mean in a pre/post study, items with low reliability would be more likely to show a backfire effect. In a previous meta-analysis (Swire-Thompson, DeGutis, & Lazer, 2020), we found preliminary evidence for this reliability-backfire relationship by comparing studies using single-item measures—which typically have poorer reliability (Jacoby, 1978; Peter, 1979)—with more reliable multi-item measures. Examining 31 studies and 72 dependent measures[1], we found that the proportion of backfire effects observed with single item measures was significantly greater than those found in multi-item measures. Notably, when a backfire effect was reported, 81% of these cases were with single-item measures (70% of worldview backfire effects and 100% of familiarity backfire effects), whereas only 19% of cases used multi-item measures. This suggests that measurement error could be a contributing factor, but it is important to more directly measure the contribution of reliability to the backfire effect.

The goal of the current study was to quantify the contribution of reliability by examining an array of 21 misinformation items and measuring the extent to which item reliability is associated with backfire rate across items. Furthermore, we aimed to characterize the association of worldview-related characteristics (personal importance, importance to society, and strength of belief) and familiarity-related characteristics (novelty and the illusory truth effect) with the backfire rate. To better capture potential familiarity and worldview/backfire associations, we included items with a wide range of importance, strength of prior belief, and novelty. Finally, to ensure that the results we observed were robust and replicable, we performed two nearly identical studies using the same items with different sets of participants. Based on previous results (Swire-Thompson, DeGutis, & Lazer, 2020; Wood & Porter, 2019), we hypothesized (a) that the correction condition's backfire rate would not be greater than the test-retest control condition and (b) that reduced item reliability would significantly predict greater backfire rates.

---

[1] 50% used single item-measures and 50% used multi-item measures.

Method

Design

For both Experiment 1 and 2, we incorporated a between-groups longitudinal design where we compared a correction condition to a test-retest control condition. The test-retest control was used to measure reliability of each item and as comparison condition for potential backfire effects in the correction condition. We included a three-week retention interval, given that both familiarity and worldview influences are theoretically more likely to occur after a long retention interval (Skurnik et al., 2005; Wittenberg & Berinsky, 2020).

Experiment 1 was identical to Experiment 2 in terms of the design, items, and retention interval, with relatively minor differences. The first difference was that Experiment 2 had a larger sample size compared to Experiment 1. Second, Experiment 1 collected the control and correction conditions in different phases and thus participants were not randomly assigned, whereas Experiment 2 randomly assigned consecutive participants to correction and control conditions. Finally, Experiment 1 randomized the item presentation order for each participant, where Experiment 2 presented the items in a random but fixed order (decided by a random number generator) to reduce order-related individual differences (Ruiz et al., 2019). Given the minor differences between the two experiments, we report both experiments together.

Sample Size Justification

The effect sizes of backfire effects vary dramatically, ranging from $\eta_p^2 = .26$ (large effect; Pluviano et al., 2019) to $\eta_p^2 = .03$ (small effect; Ecker & Ang, 2019). If we take the smallest effect size, given that our main analysis is a 2×2 between-within ANOVA, a power analysis conducted by G*Power3 (Faul et al., 2009) recommended a sample size of 108 people total to acheive a power of .95. However, since we purposively included items with a wide range of reliabilities, we substantially boosted our sample size to detect small effects for items with lower reliabilities ($\rho = \sim.3-.5$ test-retest). We therefore aimed to collect >350 participants for Experiment 1 and >500 for Experiment 2.

Participants

In Experiment 1 there were 423 crowdworkers from the U.S. recruited through Prolific Academic who took part in the pretest phase, and 388 of these participants who returned to take part in the posttest phase (92% retention). There were 175 males, 205 females, and 8 individuals choosing not to disclose their gender. Participants ranged in age between 18 and 82 ($M = 32.63$, $SD = 12.18$). There were 195 people allocated to the correction condition and 193 to the control condition. In Experiment 2 there were 600 crowdworkers from the U.S. recruited through Prolific Academic who took part in the pretest. For the posttest, 532 of these participants returned to take part (88% retention). There were 262 males, 258 females, and 12 individuals choosing not to disclose their gender. Participants ranged in age between 18 and 74 ($M = 31.35$, $SD = 11.08$). There were 268 people allocated to the correction condition and 264 to the control condition. Prolific Academic was selected to recruit participants as it is more diverse in age, race, and education than other samples of convenience such as psychology students. While our study did not select for extreme conspiratorial thinkers (such as QAnon), we note that the vast majority of studies that report eliciting the backfire effect also use similar samples of convenience rather than focusing solely on those who hold conspiratorial beliefs (see Swire-Thompson, DeGutis, & Lazer, 2020).

Stimuli

Items were created to have a broad representation across the dimensions of novelty, importance, and prior strength of belief. See Table 1 for an example of the misinformation and corrections presented, Supplementary Table 1 for all misinformation stimuli, and Supplementary Table 2 for all fact stimuli. We took both

Table 1

Example Misinformation and Correction

Type of example	Misinformation	Correction
Example of an important, familiar item	The gender pay gap is driven by women being paid less for the same job.	The gender pay gap is driven by women being paid less for the same job.
		This is false
		Data from 25 countries revealed that women earn 98% of the wages of men when doing the same job for the same employer. According to The Economist, the gender pay gap is primarily driven by the fact that women are less likely to hold high-level, high-paying jobs than men. In other words, women cluster in lower-tier jobs such as administrative roles.
Example of a less important novel item (to a US population)	"Marmite", the black salty spread from UK, is primarily meat based	"Marmite", the black salty spread from UK, is primarily meat based
		This is false
		Although some say Marmite has a meaty flavor, it is made from yeast extract. Therefore, it can be consumed by vegetarians and even vegans. In the UK, it is commonly eaten as a savory spread on toast. According to the BBC, it was invented in the late 19th Century by a German Scientist and is a byproduct of brewing beer.

societal importance and personal importance into account given that these could be distinct (Hewitt, 1976). A pilot study was conducted with 80 Prolific Academic participants from 70 items. For each item participants indicated on a 0–10 scale (a) the extent to which they believed the item, (b) whether they had heard the statement before, (c) whether it is important to the participants themselves whether it is true or false, and (d) whether it is important to society whether it is true or false. All four questions appeared at the same time on the screen. Items for the following experiments were chosen to have a distributed range of these attributes. This resulted in a final stimulus set of 21 misinformation items and 21 facts, with each claim having a corresponding correction (if misinformation) or affirmation (if true). See Supplementary Figure 1 for the histograms showing the range of familiarity, belief, personal importance and societal importance of the items included in the current study.

Several of the less important items were adapted from Swire et al. (2017), and the important items were adapted from Wood and Porter (2019), given that these were specifically selected in an attempt to inspire worldview backfire effects. The corrections repeated the initial statement, highlighted that it was false, and provided a 2–3 sentence explanation and referred to a reputable source. All corrections were designed to have similar word counts ($M = 60$, $SD = 3.08$) and Flesch-Kincaid reading levels ($M = 9.95$, $SD = .89$). Flesch-Kincaid reading levels are equivalent to U.S. grade level of education, showing the education required to read the text, as measured by Microsoft Word (Zhou et al., 2017).

Procedure

In the pretest, participants were presented with the 42 claims: 21 facts and 21 misinformation items. An equal number of misinformation and facts were given so that participants would not be biased toward true or false responses in later veracity judgements (Swire-Thompson, Ecker, et al., 2020), though for the purpose of this study, we were only interested in the misinformation items. At the commencement of the study, only participants in the correction group received instructions indicating that they would be told whether each statement was true or false. For each claim participants indicated on a 0–10 scale (a) the extent to which they believed each item, (b) whether they had heard the statement before, (c) whether it is important to *themselves* if it were true or false, and (d) whether it is important to *society* if it were true or false. In an identical fashion to the pilot study, all four questions appeared at the same time on the screen. Directly after these four questions, participants in the correction group were presented with a correction if the item was false and an affirmation if the item was true. To ensure that participants read the correction and affirmation, they were asked a final question of how surprised they were that the item was true or false.[2] The posttest for all participants was three weeks later, where the 42 items were presented again, and participants rated each item on a 0–10 scale for how much they believed them to be true. For the following results, we note that we have reported all measures, conditions, and data exclusions.

Results

We first sought to determine if our correction and control groups were well balanced across variables in both experiments. In Experiment 1 there were no significant differences between the

correction and control condition in age ($p = .377$), education ($p = .974$), gender ($p = .999$), or partisanship ($p = .134$). For Experiment 2 there were also no significant differences between the correction and control condition for age ($p = .516$), gender ($p = .081$), education ($p = .115$), or partisanship ($p = .969$).

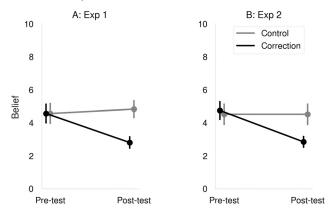### Comparing Group Level Belief Between Control and Correction Groups

To compare overall belief differences between correction and control conditions, we conducted a 2×2 between-within ANOVA on belief ratings collapsed across items, with factors fact-check (correction versus no-correction control) and pre/post retention interval. As can be seen from Figure 1, a significant interaction revealed that on average, participants in the correction condition updated their beliefs in the intended direction to a greater extent than in the control condition; Experiment 1: $F(1, 386) = 325.01$; $p < .001$; $MSE = .61$; $\eta_p^2 = .46$; Experiment 2: $F(1, 530) = 329.93$; $p < .001$; $MSE = .72$; $\eta_p^2 = .46$. As expected, when focusing on the pre versus post belief in the control condition, planned comparisons revealed a small but significant illusory truth effect, as can be seen from Figure 1 Experiment 1: $t(192) = -4.72$; $p < .001$; Cohen's $d = .22$; experiment 2: $t(263) = -5.71$; $p < .001$; Cohen's $d = .25$.

We next individually examined all 21 misinformation items at the group level with their pretest item means plotted against the posttest item means, see Figure 2. See Supplementary Figure 2 to view the fact items and corresponding item numbers and Supplemental Analyses 1 for an analysis of the fact items. The backfire zone in Figure 2 is highlighted in gray, where items would be located for them to be classified as backfiring at the group level. There were two items in the correction condition that numerically crossed into the backfire zone, "Testosterone treatment helps older men retain their memory" (observed in Experiment 1 and 2) and "Marmite, the black salty spread, is primarily meat based" (observed in Experiment 2 only). However, pre- and postmeasures did not significantly differ for either the testosterone item ($p = .935$ and $p = .437$, respectively) nor the marmite item ($p = .380$). As can be seen from Figure 2, items varied considerably in their proximity to the backfire zone, with those with initially lower beliefs closer to the backfire threshold.

One concern is that the test-retest condition is an imperfect control given that it could be subject to the illusory truth effect, where information repetition leads to increased familiarity and thus is more likely to be assumed to be true. In other words, comparing the pre/post correction condition versus the test-retest control condition may simply be measuring the difference between the illusory truth effect and the familiarity backfire effect. Indeed, as reported above, we found a small but significant increase in average misinformation belief in the

---

[2] Theoretically there is rationale to predict that surprise could decrease the likelihood of backfire effect given the *hypercorrection effect*. This is where inaccurate information is updated more successfully if an individual is more surprised when it turns out to be false (Butterfield & Metcalfe, 2001; Metcalfe, 2017). We found that in Experiment 1, surprise was not significantly correlated with backfire rate, $p = .17$. In Experiment 2, surprise was correlated with backfire rate $\rho = -.57$; $p = .008$, showing that more surprising items backfired less. To test whether surprise predicted backfire rate above and beyond reliability, we ran a hierarchical regression, entering reliability in the first model and examining the change in variance explained in the second model with both reliability and surprise. After inputting reliability, surprise did not explain additional variance in backfire rate. See Supplement 8 for full analysis.

Figure 1

Experiment 1 (Panel A) and Experiment 2 (Panel B) Belief Ratings Collapsed Across All Items, With Fact-Check (Correction Versus Control) and Pre/Posttest

Note. Error bars denote 95% confidence intervals.

control condition for both Experiments 1 and 2. To mitigate these concerns, we compared postcorrection beliefs in the correction group with first-time belief ratings in the control group (similar to Nyhan & Reifler, 2010). In other words, we examined whether correcting misinformation led to increased be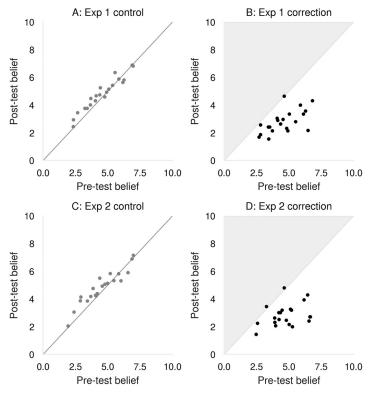lief after a three-week retention interval, in comparison with an initial misinformation belief rating in the control group. Consistent with the pre versus post results, we found that items that received a correction had significantly *lower* belief ratings three weeks later compared to pretest items in the control group: Experiment 1, $t(386) = -11.62$, $p < .001$; Cohen's $d = 1.18$; Experiment 2, $t(530) = -13.39$, $p < .001$; Cohen's $d = 1.16$).

We next compared the pretest control with the posttest correction group for each individual item (see Supplementary Table 3 for each individual result). There was only one item in one experiment where participants' beliefs were significantly greater in the posttest correction group than the pretest control group ("Marmite, the black salty spread, is primarily meat based", observed in Experiment 2; $p = .022$). However, this did not remain significant after a Holm-Bonferroni correction was applied to account for multiple comparisons.

Defining the Correction Group's Backfire Rate and the Control Group's Belief Boost Rate

Given that backfire effects have most often been reported in subgroups, we sought to go beyond comparing mean group-level beliefs by examining the number of individuals that increase belief on a particular item. We define the *backfire rate* as the percentage of people that increase their belief on a given item. Given that the control condition technically cannot

Figure 2

Pretest and Posttest Belief for Experiments 1 and 2

Note. Panels A and C show the control condition items in gray; the gray line represents the line of equality. Panels B and D show the correction condition items in black; the gray area represents the misinformation backfire zone.

backfire given that no correction was elicited, we label this as the *belief boost rate*. Though there are several other ways to measure belief change (e.g., subtraction scores, percentage change scores), we used a rate of increase because this most accurately reflects what researchers have defined as a backfire effect (Swire-Thompson, DeGutis, & Lazer, 2020). Further, this measure considerably reduced the severity of regression-to-the mean effects compared to subtraction scores (Vickers & Altman, 2001)—as demonstrated by Supplementary Figure 3. Finally, the increase rate is an "all or none" approach, which minimizes ceiling and floor effects, given that a shift from an 8 to a 10 is considered the same as a shift from a 3 to a 10.

We thus defined backfire rate and belief boost rate as the percentage of individuals that increase their belief by two or more points on the 11-point scale per item. We chose two or more points because a one-point increase might be explained by indecision between two points. For instance, if a participant holds a belief at 7.5 out of 10, they could feasibly round down to 7 at time point 1 and round up to 8 at time point 2. For robustness, we replicated all analyses using a one-point increase in belief and note any occasion where the findings are not replicated.

Comparing the Backfire Rate and Belief Boost Rate

We first conducted a *t* test comparing item backfire rate (correction condition) with the belief boost rate (control condition) using a Bayesian framework. This is because we hypothesized corrected misinformation would *not* increase more than the control condition, and Bayes factors (BFs) can quantify relative evidence favoring the null hypothesis. We thus compared the evidence in favor of the null hypothesis (that corrected misinformation will not increase belief more than the control condition) with evidence against the alternative hypothesis (that corrected misinformation will increase belief more than the control). We report our BF as favoring the null hypothesis (BF_{01}), and conducted the analysis in JASP (JASP Team, 2020).

For Experiment 1, we found strong evidence in favor of the null hypothesis, $BF_{01} = 78.76$, suggesting that corrected misinformation did not increase belief more than the control condition. For consistency, we also conducted a *t* test using null hypothesis testing and found that the control condition has a significantly higher rate of belief increase than the correction condition, $t(20) = -9.29$; $p < .001$. In fact, the percentage of people in the control condition that increased their belief ($M = 24.03\%$) was nearly twice the rate of those in the correction condition ($M = 12.65\%$).

Experiment 2 replicated these findings. We first conducted a Bayesian *t* test and found evidence in favor of the null hypothesis, $BF_{01} = 73.27$, again suggesting that corrected misinformation did not increase belief more than the control condition. Indeed, a significantly higher percentage of participants in the control condition increased their belief ($M = 14.05\%$) than those in the correction condition ($M = 24.82\%$), $t(20) = -8.73$; $p < .001$.

We next tested whether the backfire rate was greater on any individual item in comparison to the control condition's belief boost rate. As can be seen from Figure 3, the backfire rate varied dramatically across items, ranging from 4.62% to 32.31% in the correction condition for Experiment 1 and from 2.99% to 33.96% in Experiment 2. Using the z-ratio to test whether two independent

proportions are significantly different, we found that the backfire rate was never significantly greater than the control condition's belief boost rate for any item. See Supplementary Table 3 for all analyses. Note that all the significant effects in Supplementary Table 3 were from the correction condition having a significantly lower backfire rate than the control condition. Together, these results suggest that correcting misinformation does not increase belief at the group or item level; this remains the case for both group means and backfire rates when compared to either a test-retest no-correction control or pretest belief.

Attributes Associated With Backfire Effects

We next sought to examine the best predictors of backfire effects in the correction condition. Although no items in the correction condition showed greater backfire rates compared to the control condition's rate of increase, there are still mechanistic and practical reasons to further examine backfire rates in the correction condition alone. Mechanistically, it is crucial to know if backfire effects after a correction are more related to measurement error (i.e., reliability) or rather theoretically meaningful factors such as importance, strength of belief, and novelty, or the illusory truth effect. Practically, it is important for fact-checkers to know if there are certain types of items that are particularly resistant to correction.

Reliability

To robustly estimate the test–retest reliability of each item, we correlated pre versus posttest belief ratings from participants who were in the control group collapsed across both experiments. We used Spearman's ρ given that the data were not normally distributed. As can be seen in Figure 4, reliability was strongly correlated with the backfire rate in both Experiment 1 ($\rho = -.61$, $p = .004$) and Experiment 2 ($\rho = -.73$, $p < .001$). In other words, 37% and 53% of the variance in the backfire rate was explained by test–retest reliability in Experiment 1 and 2, respectively[3].
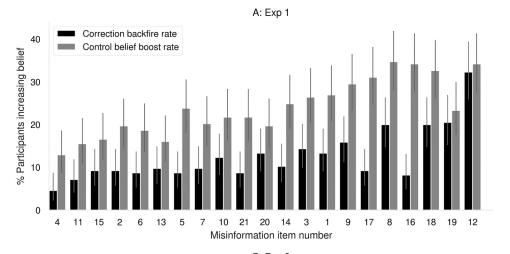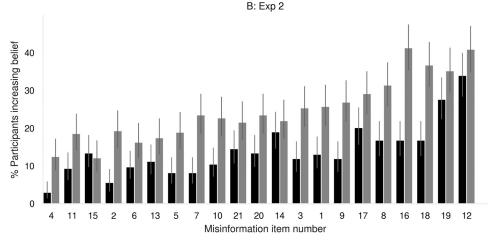
Worldview and Familiarity

We next tested whether theoretically driven factors can explain the large item-to-item variation in backfire rate. For the worldview backfire effect model, we correlated backfire rates with participants' ratings of societal importance, personal importance, and strength of belief. As can be seen in Table 2, the correlations of backfire rate with societal and personal importance were both nonsignificant in both experiments. There was a weak nonsignificant negative relationship between backfire rate and strength of belief in Experiment 1, and in Experiment 2 there was a significant negative relationship. In other words, the *lower* the participants' precorrection belief, the more likely the item was to backfire. This is inconsistent with the worldview backfire effect hypothesis predicting stronger beliefs are associated with greater propensity to backfire and is more consistent with a regression-to-the-mean effect.

---

[3] In the current paper we focus on backfire effects rather than the extent to which misinformation is corrected. However, it is important to note that reliability was not correlated with posttest belief in either Experiment 1, $p = .25$ or Experiment 2, $p = .65$.

Figure 3

Variation of Backfire Rate for Experiments 1 and 2 on Each Item in Control and Correction Conditions

Note. Ordered from lowest to highest mean backfire rate. Error bars denote 95% confidence intervals.

For the familiarity backfire effect model, we correlated backfire rate with participants' novelty ratings and illusory truth scores. Theoretically, items which are more likely to increase due to the illusory truth effect in the control condition should also be more likely to increase in the correction condition. We calculated the propensity for items to increase belief due to the illusory truth effect by subtracting the posttest belief ratings from pretest belief ratings in the control group. For robustness, we also conducted all analyses using the increased belief rate approach (the percentage of individuals that increase their belief by two or more points), and all findings were replicated.

Experiment 1 found a marginal correlation between backfire rate and novelty, and Experiment 2 revealed a significant correlation[4]. These correlations are negative, reflecting that items that are less familiar (and more novel) were more likely to backfire. This provides initial support for the notion that novel items are more likely to backfire. We also found that illusory truth scores were significantly correlated with backfire rate in both Experiment 1 and 2, suggesting that the same mechanisms that produce the illusory truth effect may produce the familiarity
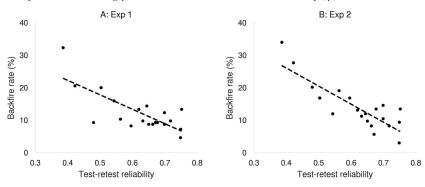
backfire effect. However, it is imperative to check that these associations are still present after accounting for item reliability.

Do Theoretically Hypothesized Factors Predict Backfire Effects Beyond Reliability?

Given that reliability, novelty, and illusory truth all significantly predicted backfire rates, we next sought to examine whether they predicted unique variance. We did not consider societal importance, personal importance, nor strength of belief since these associations were nonsignificant or significant in the opposite direction hypothesized. The following analyses were exploratory due to the low number of items in relation to the number of predictors (Harrell, 2015). We examined reliability and novelty, and reliability and illusory truth in separate models due to the low number of items.

---

[4] Note that in Experiment 1 novelty becomes significant if we use a backfire increase threshold of 1 point ($\rho = -0.55$; $p = .01$), rather than our standard increase threshold of 2 points.

Figure 4

Backfire Rate Is Strongly Correlated With Test-Retest Reliability (ρ)

To clarify if item novelty predicted backfire rate above and beyond reliability, we ran a hierarchical regression, entering reliability in the first model and examining the change in variance explained in the second model with both reliability and novelty. In Experiment 1, we found that after inputting reliability, novelty did not explain additional variance in backfire rate (see Supplementary Table 4 for the full analyses). In Experiment 2, we found that even after accounting for reliability, novelty continued to independently predict backfire rate. However, it only explained an additional 6.58% of variance in contrast to the 71.04% that reliability explained (see Supplementary Table 5). Also, these findings must be interpreted with caution, given that novelty and reliability were themselves highly correlated at $\rho = .76$, $p < .001$, and multicollinearity in a regression model makes it difficult to distinguish between the individual effects (Bergtold et al., 2018). Future research should investigate whether novelty explains the backfire rate beyond reliability using a greater number of items and with more of a dissociation between reliability and novelty.

Finally, we ran a hierarchical regression to clarify whether illusory truth score predicted the backfire rate above and beyond item reliability. We entered reliability in the first model and examined the change in variance explained in the second model with both reliability and the illusory truth score. In both Experiment 1 and Experiment 2 we found that after inputting reliability, the illusory truth score failed to explain additional variance in backfire rate (see Supplementary Tables 6 and 7 for full analyses). However, it is important to note that illusory truth was also highly correlated with reliability in Experiment 1, $\rho = -.69$, $p < .001$, and Experiment 2, $\rho = -.68$, $p < .001$.

Discussion

The current study conducted a detailed examination of the backfire effect by using a rigorous longitudinal design and considering item reliability, approaches commonly employed in the clinical intervention field but rarely used in misinformation studies. We conducted two nearly identical experiments in different samples as a check of the robustness and replicability of the findings. Our results demonstrate that corrections reduce belief in misinformation compared to test-retest and initial belief control conditions at the group level, both overall and for individual items. This is consistent with recent studies (Wood & Porter, 2019) and reinforces the use and potential power of corrections. It suggests that fact-checkers should never avoid giving a correction due to backfire concerns, consistent with the recommendations from the recent Debunking Handbook (Lewandowsky et al., 2020). Notably, we also found large variations in individual item backfire rates. This variability may explain why some studies have elicited the backfire effect while others have not (Swire-Thompson, DeGutis, & Lazer, 2020).

A key finding of the current study is that the backfire rate of each corrected item was strongly predicted by its test–retest reliability. In other words, less reliable items backfired at a substantially higher rate than more reliable items. The effect size of this association was large for both Experiments 1 and 2, explaining 36% and 53% of variance in backfire rates, respectively. This is consistent with previous work showing that backfire effects are more often elicited with less reliable single-item measures compared to more reliable multi-item measures (Swire-Thompson,

Table 2

Spearman Correlations Between Backfire Rate and Worldview and Familiarity-Related Attributes, ρ and p Values Shown

	Worldview-related attributes			Familiarity-related attributes	
Measure	Importance (society)	Importance (personal)	Pretest belief	Familiarity (novelty)	Illusory truth
Backfire rate	−.15	−.23	−.23	−.40	.58*
(Exp 1)	$p = .513$	$p = .314$	$p = .320$	$p = .081$	$p = .006$
Backfire rate	−.11	−.19	−.53*	−.82*	.59*
(Exp 2)	$p = .639$	$p = .419$	$p < .013$	$p < .001$	$p = .005$

Note. Asterisk denotes significance $p < .05$ after a Holm-Bonferroni correction has been applied.

DeGutis, & Lazer, 2020). While there have been previous papers highlighting the nonreplicable nature of backfire effects (e.g., Wood & Porter, 2019), the current findings provide a potential mechanism for this poor replicability. There are two reasons why reliability might be low: (a) measurement error or (b) people may not know where they stand on certain topics, and their beliefs may not be well formulated. As discussed in Swire-Thompson, DeGutis, & Lazer (2020), future studies could tease apart measurement error from inconsistency of beliefs on a topic by using several items to measure participants' within-belief consistency (e.g., higher consistency would indicate better formulated beliefs) as well as explicitly asking participants to rate how well-formed or they perceive their beliefs to be.

Beyond reliability, there was no evidence that worldview-related item attributes were related to backfire rates. Items which were perceived to be more important, either to one's self or to society, were not more likely to backfire. Similarly, there was no evidence that the items that were believed strongly prior to the correction were more likely to backfire. These results highlight the importance of explicitly testing the underlying mechanisms for why backfire effects occur. While it does not necessarily mean that worldview backfire effects never occur, they are unlikely to be driven by perceived importance or strength of belief, as some researchers have suggested (Flynn et al., 2017; Lewandowsky et al., 2012; Nyhan & Reifler, 2010; Nyhan & Reifler, 2015). Future research should investigate other assumptions that are thought to motivate the worldview backfire effect, such as internal counterarguing (Nyhan & Reifler, 2010), strength of identity associated with claim (Lewandowsky et al., 2012), or when a correction is perceived to be acutely threatening (Ecker et al., 2021). It will be important to determine if these effects explain backfire rates above and beyond reliability.

The familiarity-related item attributes were initially promising explanations of our observed backfire rates in the correction condition, with both novelty and illusory truth effect being significantly correlated with backfire rate. However, the illusory truth score did not explain any unique variance once reliability was taken into account. Novelty did explain a small amount of unique variance above reliability, but this only occurred in Experiment 2, and the findings must be interpreted cautiously given the multicollinearity between novelty and reliability. Potential explanations for the association between novelty and backfire rate could reflect that individuals who are presented with novel items are unsure where they stand on the subject and thus answer more randomly.

The current findings have important implications for future research and fact-checking. Future research should investigate other psychological phenomenon relevant to the processing of misinformation in the context of reliability. In particular, the illusory truth effect should be examined given that it was shown to be more likely to be elicited with low-reliability items. More generally, our results suggest that it is crucial for future research into misinformation to (a) use reliable measures and when possible boost reliability by using multi-item measures, (b) report the reliability of their measures, and (c) take reliability into account in analyses. Considering reliability could also help the broader field of misinformation and produce more replicable findings. For instance, knowing measurement reliability could allow researchers to put greater weight into effects observed with highly reliable measures, avoid overinterpreting measures with low reliability, and prevent researchers from planning new studies based on noisy measures.

With regard to real-world applications, these findings suggest that fact-checkers should not avoid giving corrections to any groups due to concerns surrounding the worldview backfire effect, given that we still do not know what motivates it and indeed it may not occur at all. Additionally, fact-checkers should not avoid giving corrective information to groups that they assume will find the information important or believe the information strongly. Fact-checkers may wish to avoid correcting misconceptions that are largely unknown, given that people are likely to have no firm belief on the matter, so time will be better spent on different fact-checks.

A limitation of the current study is that we did not have sufficient items where novelty and reliability were dissociated. Additionally, there may be some constraints on the generalizability of our findings given our participants and materials (Simons et al., 2017). For instance, our population was slightly more educated and more liberal than the general U.S. population (see Berinsky et al., 2012). Furthermore, our experimental setting differed from real-world scenarios such as fact-checking websites or social media platforms. It would be important to validate the current findings in these real-world settings, potentially targeting theoretically favorable groups such as those with strong antivaccination sentiments. However, our sample was relatively large, had a good age range, and included a wide array of items, suggesting some degree of generalizability of the current findings.

In sum, the current study uses a powerful framework to examine belief change and the correction of misinformation. The results provide unequivocal support for the benefits of correcting misinformation and suggest that backfire effects are driven substantially more by measurement error or inconsistencies in beliefs rather than the psychological mechanisms proposed to explain them. The contemporary study of misinformation continues to advance the understanding of issues vital to fostering informed citizenries and democracies, and the accumulating evidence is that media and information workers and their institutions should not hesitate to intervene and correct misinformation.

Context of the Research

The idea for this study originated when writing Searching for the Backfire Effect: Challenges and Recommendations (Swire-Thompson, DeGutis, & Lazer, 2020) and was conceptualized by Drs. Swire-Thompson and DeGutis. While reviewing the backfire literature, we found that the backfire effect was more likely to occur in studies using single item measures, which are often less reliable, in comparison to multi-item measures. Based on these findings, we sought to more directly assess how item reliability contributes to the likelihood of observing a backfire effect. The current work is related to the research program of Dr. Swire-Thompson as she investigates why certain individuals are predisposed to refrain from belief change even in the face of good corrective evidence and how corrections can be designed to maximize impact. It is related to Dr. DeGutis's interests in the contribution of measurement reliability to individual differences and effects of interventions. The current study is relevant to Dr. Wihbey's research which focuses on news and social media from a journalistic perspective, and to Dr. Lazer who investigates misinformation online. This work can be extended in the future by investigating other psychological phenomena relevant to misinformation and accounting for reliability.

References

Alter, A. L., & Oppenheimer, D. M. (2006). Predicting short-term stock fluctuations by using processing fluency. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(24), 9369–9372. https://doi.org/10.1073/pnas.0601071103

Autry, K. S., & Duarte, S. E. (2021). Correcting the unknown: Negated corrections may increase belief in misinformation. *Applied Cognitive Psychology*, *35*, 960–975. https://doi.org/10.1002/acp.3823

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, *121*(4), 446–458. https://doi.org/10.1037/0096-3445.121.4.446

Bergtold, J. S., Yeager, E. A., & Featherstone, A. M. (2018). Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. *Journal of Applied Statistics*, *45*(3), 528–546. https://doi.org/10.1080/02664763.2017.1282441

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491–1494. https://doi.org/10.1037/0278-7393.27.6.1491

Cameron, K. A., Roloff, M. E., Friesema, E. M., Brown, T., Jovanovic, B. D., Hauber, S., & Baker, D. W. (2013). Patient knowledge and recall of health information following exposure to "facts and myths" message format variations. *Patient Education and Counseling*, *92*(3), 381–387. https://doi.org/10.1016/j.pec.2013.06.017

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

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

Ecker, U. K., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology*, *40*(2), 241–260. https://doi.org/10.1111/pops.12494

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 & Cognition*, *6*(2), 185–192. https://doi.org/10.1016/j.jarmac.2017.01.014

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

Ecker, U. K., Sze, B. K., & Andreotta, M. (2021). Corrections of political misinformation: no evidence for an effect of partisan worldview in a U.S. convenience sample. *Philosophical Transactions of the Royal Society B*, *376*(1822), Article 20200145. https://doi.org/10.1098/rstb.2020.0145

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*(5), 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

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. https://doi.org/10.1111/pops.12394

Guess, A., & Coppock, A. (2020). Does counter-attitudinal information cause backlash? Results from Three Large Survey Experiments. *British Journal of Political Science, 50(4)*, 1497–1515. https://doi.org/10.1017/S0007123418000327

Haglin, K. (2017). The limitations of the backfire effect. *Research & Politics*. Advance online publication. https://doi.org/10.1177/2053168017716547

Harrell, F. E. Jr. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer. https://doi.org/10.1007/978-3-319-19425-7

Hewitt, J. P. (1976). *Self and society. A symbolic interactionist social psychology.* Allyn & Bacon.

Horne, Z., Powell, D., Hummel, J. E., & Holyoak, K. J. (2015). Countering antivaccination attitudes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(33), 10321–10324. https://doi.org/10.1073/pnas.1504019112

Jacoby, J. (1978). Consumer research: How valid and useful are all our consumer behavior research findings? A state of the art review. *Journal of Marketing*, *42*(2), 87–96. https://doi.org/10.1177/002224297804200213

JASP Team. (2020). JASP (Version 0.14.1)[Computer software].

Jerit, J., & Zhao, Y. (2020). Political misinformation. *Annual Review of Political Science*, *23*(1), 77–94. https://doi.org/10.1146/annurev-polisci-050718-032814

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M., Swire-Thompson, B., van der Linden, S., . . . Zaragoza, M. S. (2020). The debunking handbook 2020. https://doi.org/10.17910/b7.1182

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

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. *Psychological Science*, *11*(5), 424–428. https://doi.org/10.1111/1467-9280.00282

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465–489. https://doi.org/10.1146/annurev-psych-010416-044022

Nyhan, B. (2021). Why the backfire effect does not explain the durability of political misperceptions. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(15), Article e1912440117. https://doi.org/10.1073/pnas.1912440117

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

Nyhan, B., & Reifler, J. (2015). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine*, *33*(3), 459–464. https://doi.org/10.1016/j.vaccine.2014.11.017

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

Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, *16*(1), 6–17. https://doi.org/10.1177/002224377901600102

Pluviano, S., Watt, C., Ragazzini, G., & Della Sala, S. (2019). Parents' beliefs in misinformation about vaccines are strengthened by pro-vaccine campaigns. *Cognitive Processing*, *20*(3), 325–331. https://doi.org/10.1007/s10339-019-00919-w

Ruiz, S., Chen, X., Rebuschat, P., & Meurers, D. (2019). Measuring individual differences in cognitive abilities in the lab and on the web. *PLoS*

ONE, *14*(12), Article e0226217. https://doi.org/10.1371/journal.pone.0226217

Schmid, P., & Betsch, C. (2019). Effective strategies for rebutting science denialism in public discussions. *Nature Human Behaviour*, *3*(9), 931–939. https://doi.org/10.1038/s41562-019-0632-4

Schwarz, N., & Jalbert, M. (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. In R. Greifeneder, M. Jaffé, E. J. Newman & N., Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation* (pp. 73–89). Routledge.

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. *Advances in Experimental Social Psychology*. Advance online publication. https://doi.org/10.1016/S0065-2601(06)39003-X

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128.

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. https://doi.org/10.1086/426605

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science*, *4*(3), Article 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*(12), 1948–1961. https://doi.org/10.1037/xlm0000422

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory & Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Swire-Thompson, B., Ecker, U. K., Lewandowsky, S., & Berinsky, A. J. (2020). They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology*, *41*(1), 21–34. https://doi.org/10.1111/pops.12586

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

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

Vickers, A. J., & Altman, D. G. (2001). Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ (Clinical Research Ed.)*, *323*(7321), 1123–1124. https://doi.org/10.1136/bmj.323.7321.1123

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*(4), 401–422. https://doi.org/10.1093/ijpor/edu005

Wittenberg, C., & Berinsky, A. J. (2020). Misinformation and its correction. N. Persily & J. A. Tucker, (Eds.), *Social Media and Democracy: The State of the Field, Prospects for Reform* (pp. 163–198). Cambridge University Press.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Zhou, S., Jeong, H., & Green, P. A. (2017). How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, *60*(1), 97–111. https://doi.org/10.1109/TPC.2016.2635720

Zuo, X. N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*, *3*(8), 768–771. https://doi.org/10.1038/s41562-019-0655-x