Twitter's Disputed Tags May Be Ineffective at Reducing Belief in Fake News and Only Reduce Intentions to Share Fake News Among Democrats and Independents

Jeffrey Lees, Abigail McCarter, Dawn M. Sarno

Abstract. Throughout the 2020 US elections, one of Twitter's defenses against misinformation was its "This claim has been disputed" tags. The utility of such tags, however, remains unclear. A survey-based experiment, meant to simulate the Twitter environment, with a convenience sample of 318 US participants found that while disputed tags reduced the sharing of misinformation among Democrats and Independents, they had no effect on the sharing habits of Republicans and did not reduce belief in fake news for any group. We also found that higher scores on the Cognitive Reflection Test (a measure of analytical rather than intuitive thinking) correlated with lower belief in fake news, but had no relationship with sharing habits. Further, conservatism positively correlated with belief in and sharing intentions for tagged false headlines, but not untagged false headlines or true headlines. Our results suggest that the tags employed by Twitter to combat the spread of fake news may have been ineffective at reducing belief in fake news, and may only have attenuated fake news sharing among Democrats and Independents.

1 Introduction

The spread of misinformation about political (Jerit and Zhao 2020), scientific (Maertens, Anseel, and Linden 2020), and medical (Romer and Jamieson 2020) issues has caused serious concerns among the public, policymakers, and scientists alike. The rapid proliferation of misinformation and fake news (false news headlines presented as legitimate) on social media platforms (Cinelli et al. 2020; Pennycook and Rand 2021) has led many to argue that social media companies bear some responsibility for reducing the spread of misinformation on their platforms. For example, both Facebook (Guynn 2017) and Twitter (Ortutay 2021; Roth and Pickles 2020) implemented, then ceased, using tags like "This claim about election fraud is disputed" or "This is disputed" (i.e.,"disputed" tags) to identify posts and headlines deemed false or misleading (see Figure 1 on the following page). Facebook cited a lack of effectiveness of "disputed" tags and concerns over "backfire" effects as reasons for shifting their misinformation mitiga-

doi:10.54501/jots.v1i3.39

tion efforts (Smith 2017), despite ample research suggesting that fact-checking and other lightweight interventions reduce the sharing of misinformation (Lewandowsky and Linden 2021; Pennycook, McPhetres, et al. 2020; Walter et al. 2020), and little evidence that backfire effects from fact-checking are real (Nyhan 2021; Swire-Thompson, DeGutis, and Lazer 2020; Wood and Porter 2019). Twitter cited the need for "more context" as a reason for no longer using "disputed" tags (Twitter Support 2021), choosing instead to label them with ostensibly clearer labels such as "misleading."

Figure 1: Example disputed tags from Facebook, Twitter, and the present study.

A dominant view among psychological scientists is that a *lack of attentiveness*, not political bias, is what drives the sharing of misinformation on social media platforms (Epstein et al. 2021; Pennycook and Rand 2021, 2019b). According to this account, partisans are able to discern false from true information independent of whether it is concordant or discordant with their ideology, yet a lack of attentiveness when it comes to *sharing* behaviors leads people to inadvertently share information they would correctly identify as false otherwise. Since people are motivated to be accurate in their social judgments (Pennycook, Epstein, et al. 2021), this account suggests that nudging people out of their inattentiveness and toward a reflective mindset will reduce the sharing of misinformation. For example, a recent large-scale experiment performed on Twitter found that priming Twitter users with an accuracy-mindset nudge reduced the sharing of political misinformation (Pennycook, Epstein, et al. 2021).

Given the evidence for the effectiveness of accuracy nudges, and that disputed tag warnings are enough to modestly reduce misinformation sharing on Facebook (Clayton et al. 2020), one could reasonably hypothesize that Twitter's retired "disputed" tags successfully increased attentiveness and reduced misinformation sharing when it was deployed on Twitter. However, scholars have noted that these interventions are not equally effective across parties. Experiments examining accuracy nudges tend to find significantly lower levels of truth discernment among Republicans compared to Democrats (Gawronski 2021). Moreover, a reanalysis of COVID-19-based accuracy nudge data from Pennycook, McPhetres, et al. (2020) suggests that the effects are driven largely by Democrats, and that the nudges are least effective on Republicans (see Letter to the Editor in Pennycook, McPhetres, et al. (2021)), although the original authors have disputed this claim directly (see response to Letter to the Editor in Pennycook and Rand (2022)). These debates parallel larger bodies of evidence suggesting significant asymmetries in

partisans' fake news belief and consumption, where conservatives are generally more likely to believe and share fake news and misinformation relative to liberals (Garrett and Bond 2021; Jost et al. 2018; Pereira, Harris, and Bavel 2021)).

In the present work we seek to examine the effectiveness of Twitter's "disputed tags" in reducing belief in and sharing of fake news headlines. In an online survey we asked Democrats, Independents, and Republicans to respond to a series of headlines in tweet form, which they viewed in random order. Ten of the headlines were true, 10 of the headlines were false and had a "this claim is disputed" tag, and 10 were untagged false headlines. For each headline, we asked participants if they would be willing to share the headline and whether they believed the headline was accurate. We also measured participants' levels of cognitive reflectiveness, impulsivity, and ideology. We included an impulsivity measure because users who demonstrate a lack of cognitive reflectiveness are likely to also be impulsive, as this relationship has been demonstrated in the phishing domain (Kumaraguru et al. 2007). We hypothesized that the disputed tags would be effective at reducing belief in and the willingness to share fake news, relative to untagged fake news, and that these effects would be greater for unreflective and impulsive participants.

Contrary to our hypotheses, there was no difference in the perceived accuracy of untagged and tagged false headlines, suggesting that the presence of a "disputed" tag on the headline did not reduce belief in the accuracy of the headline. We did observe an effect of the tags on sharing intentions, such that participants were less likely to say they would share tagged versus untagged headlines. However, this effect only appeared for Democrats and Independents, but not for Republicans. We also found that participants' ideology, but not their levels of cognitive reflectiveness, interacted with the effect of the tags. Participants who were more conservative were more likely to share and believe false headlines with tags—relative to liberal participants—but no such correlation with ideology arose for true untagged headlines or false untagged headlines.

2 Method

2.1 Open Science

All materials, data, and analysis scripts needed to replicate this study are available online (https://osf.io/h65nv/). Data collection was preregistered (https://osf.io/vj35r/), as was a series of analyses testing confirmatory hypotheses. However, many of the results presented in this paper are from non-preregistered analyses, including all findings related to party identification and ideology. Details on the preregistered and non-preregistered analyses can be found in Section 2.6 on page 6.

2.2 Participants

Data were collected June 11–15, 2021, using the Prolific survey platform (Palan and Schitter 2018). Using Prolific's prescreen functions, we collected a convenience sample with an equal number of self-identified Democrats, Republicans, and Independents. Participants were not prescreened based on usage of Twitter, and the base rate of willingness to share in the experiment was 16% (the base rate of willingness to share these and other headlines from Pennycook, Epstein, et al. (2021) varied 15–35%). We collected 328 responses, and 10 participants failed the attention check, leaving a final N = 318; Mage = 33.3, 34.6% Democrat, 32.7% Republican, 32.7% Independent, 49.7% female, 47.2% male, 3.1% non-binary/unlisted, 73.0% White. Self-reported ideology (1–10 Likert, "Very Liberal" – "Very Conservative) was M = 4.35, SD = 3.13, indicating a

slight liberal skew but overall well distributed across the ideological spectrum.

2.3 Statistical Power

We conducted Monte Carlo simulations, using the simr R package (Green and MacLeod 2016), of the regression model testing the standardized effect of the False Tagged condition relative to the False Untagged condition on sharing intention, as this is of primary interest in the present research. The simulations were performed on the exact model reported below in the results, namely the main effects model without controls or interactions for party identification. Each sensitivity estimate was based on 1,000 simulations and an alpha = 0.05. We found that the sharing intentions model was sensitive enough to detect a standardized effect of β = -0.18 with 81.2% power, 95% CI = [76.6%, 83.6%], and sensitive enough to detect a standardized effect of β = -0.21 with 89.7% power, 95% CI = [87.6%, 91.5%].

2.4 Procedure

This study utilized a 3x1 repeated measures within-subjects design, where each participant was exposed to all experimental conditions (tagged false, untagged false, and true headlines), and for each headline asked the same set of questions (i.e., repeated measures). Judd, Westfall, and Kenny (2017) provide a useful typology for understanding this design, which delineates participants, targets (here, headlines) and condition (see Table 2 therein). First, participants and condition are crossed (C), meaning participants are exposed to the tagged false, untagged false, and true headlines conditions. Second, headlines are nested (N) within condition, meaning that each unique headline is fixed to a condition. Third, participants and headlines are crossed (C), meaning that participants see all headlines. Judd, Westfall, and Kenny (2017) refer to this as a "CNC" design.

After providing informed consent and reading basic instructions, all participants responded to a combined 7-item numerical and non-numerical Cognitive Reflection Test (CRT) (Frederick 2005; Thomson and Oppenheimer 2016) and Version 11 of the 30-item Barrett Impulsivity Scale (BIS) (Patton, Stanford, and Barratt 1995) to measure participants' levels of reflective thinking and impulsivity. Both of these measures assess an individual's tendency to think critically about information and avoid rash decisions. CRT and BIS-11 order was counterbalanced and item ordering was randomized. Participants then responded to the attention check "Please, in the box below, write out the answer to the following math problem, capitalizing the first letter of your answer (e.g. 'Eight,' not 'eight' or '8'). What does one plus three equal?"

Afterward, participants read and rated the perceived reliability of 30 headlines in Tweet format, in randomized order. Ten headlines were true and untagged, 10 headlines were false and untagged, and 10 headlines were false and tagged "(!) This claim has been 'disputed'" in accordance with Twitter's tags (see Figure 2 on the next page). Note "This claim has been disputed" is a genericized version of Twitter's disputed tags, which varied widely in their real-world implementation (i.e., Twitter's tags sometimes directly referenced the topic at hand, and were sometimes in the present tense).

Our 30 true and false headlines (see Appendix) were a randomly selected subset of a larger corpus of left-right balanced 76 true and 70 false political headlines from Pennycook, Binnendyk, et al. (2021). We generated our stimuli using Tweetgen (https://www.tweetgen.com/create/tweet.html). We then randomly determined which 10 of the 20 false headlines we would tag. For each headline participants were asked "To the best of your knowledge, how accurate is the claim in the above headline?" (1–6

Likert, “Extremely Inaccurate” – “Extremely Accurate”) and “If you were to see the above article on social media, how likely would you be to share it? (1–6 Likert, “Extremely Unlikely” – “Extremely Likely”), both adopted from Pennycook, Binnendyk, et al. (2021).

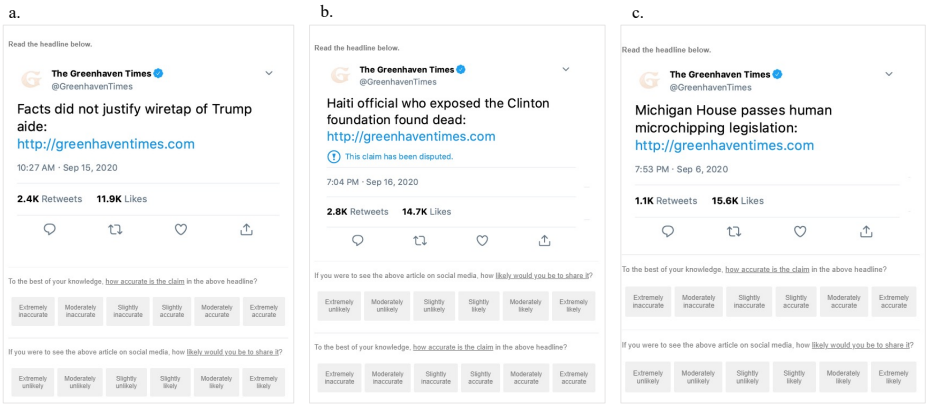

Figure 2: Example a. True headline, b. Tagged False Headline, and c. Untagged False Headline, all with the accompanying measures of perceived accuracy and sharing intentions.

After responding to all 30 headlines, participants provided basic demographic information and proceeded to a debrief slide. In the debrief, participants were informed that many of the headlines they saw were false, and participants were shown the images of the true headlines. Participants had the option to also view the false headlines. 30.5% of participants opted to view the false headlines before leaving the survey. By party, the proportion of participants who chose to view the false headlines was 39.1% of Democrats, 30.8% of Independents, and 21.2% of Republicans, and the pairwise contrast between Democrats and Republicans was statistically significant, ptukey = 0.012. Viewing the false headlines was not associated with sharing intentions of the false (tagged and untagged) headlines (b = 0.02, p = 0.024), but was positively associated with belief in the false headlines (b = 0.19, p = 0.029), meaning that participants who were more likely to believe fake news were also more likely to view the list of false headlines in the debrief.

2.5 Analyses

All analyses were conducted using restricted maximum likelihood linear mixed-effects modeling with crossed random intercepts for participants and headline to model the hierarchical dependencies within the data, as all judgments are nested within-participant and nested within-headline. We take a maximal random structure approach to modeling random slopes, and per recent guidelines (Barr et al. 2013; Brauer and Curtin 2018)(Barr et al., 2013; Brauer and Curtin, 2018) we model random slopes for all within-unit predictors, namely random slopes for condition within participant, and random slopes for individual difference measures within stimuli. If models would not converge under any available optimizers then we began removing secondary random slopes. All p-values were derived through Welch-Satterthwaite degrees of freedom approximation using the lmerTest R package (Kuznetsova, Brockhoff, and Christensen 2017), and all post-hoc tests utilized the Tukey method for p-value correction. Perceived accuracy, sharing intention, CRT, and BIS-11 responses were z-scored to allow for comparisons of standardized effects. The fixed effect for within-subjects Condition (True, False Un-

tagged, False Tagged) and the fixed effect for party identification (Republican, Democrat, Independent) were sum coded.

2.6 Preregistered Hypotheses and Non-preregistered Analyses

All hypotheses were confirmatory and preregistered (https://osf.io/vj35r). Hypotheses 1a–b predicted that perceived accuracy (1a) and willingness to share (1b) would be highest for True Headlines and lowest for Tagged False Headlines, and Non-Tagged False Headlines would be in between, with all contrasts being significantly different. Hypotheses 2a–b predicted that the main effects predicted by H1a–b would be significantly moderated by cognitive reflectiveness, such that the difference in perceived accuracy (2a) and willingness to share (2b) between Untagged and Tagged False Headlines would be greater for participants lower in cognitive reflectiveness relative to those higher in cognitive reflectiveness. Hypotheses 3a–b predicted that main effects predicted by H1a–b would be significantly moderated by impulsivity, such that the difference in perceived accuracy (3a) and willingness to share (3b) between Untagged and Tagged False Headlines would be greater for participants higher in impulsivity relative to those lower in impulsivity.

Any and all analyses that are not stated as numbered hypotheses were not preregistered. Notably, none of the analyses below investigating party identification or ideology were preregistered. Such analyses were pursued in response to a lack of support for many of the stated hypotheses. Additionally, the preregistration erroneously neglected to specify the modeling of random intercepts for headline, and erroneously stated that participants see 15 headlines, rather than 30.

3 Results

3.1 Belief

In partial support for Hypothesis 1a, across all partisan groups we found that participants perceived both tagged (β = -0.80, $p < 0.001$) and untagged (β = -0.76, $p < 0.001$) as less accurate than true headlines; however, there was no difference between untagged and tagged false headlines (β = -0.04, $p = 0.766$). To test Hypothesis 2a we interacted participants' performance on the Cognitive Reflection Test (CRT) with condition and observed a significant interaction ($p = 0.008$). This interaction indicated that the CRT-belief slope was different by condition. We found that CRT scores negatively correlated with belief in tagged (β = -0.07, $p = 0.011$), but not associated with belief in untagged headlines (β = -0.05, $p = 0.072$) or belief in the true headlines (β = 0.05, $p = 0.057$). At no level of cognitive reflectiveness (+/-1 SD) did we observe a difference in belief between tagged and untagged fake news. Rather, the gap between belief in true and false headlines grew as participants' levels of cognitive reflectiveness increased; see Figure 3 on the next page). This overall suggests that cognitive reflectiveness's role in the effect of disputed tags was minimal. We also found no support for Hypothesis 3a, as self-reported impulsivity (BIS rating) was not associated with belief (β = 0.04, $p = 0.061$), nor did it interact with condition (all ps > 0.160). The full regression tables for these analyses, and all analyses presented below, can be found in Appendix B.

The main effect of condition on belief was qualified by two significant interactions with party identification ($p = 0.039$ and $p = 0.009$). These interactions did not indicate that the tags were effective for any political subgroup; instead, they indicated that the tags induced partisan correlations within condition. All pairwise contrasts utilized Tukey-adjusted p-values. Among tagged false headlines, Republicans were more likely to per-

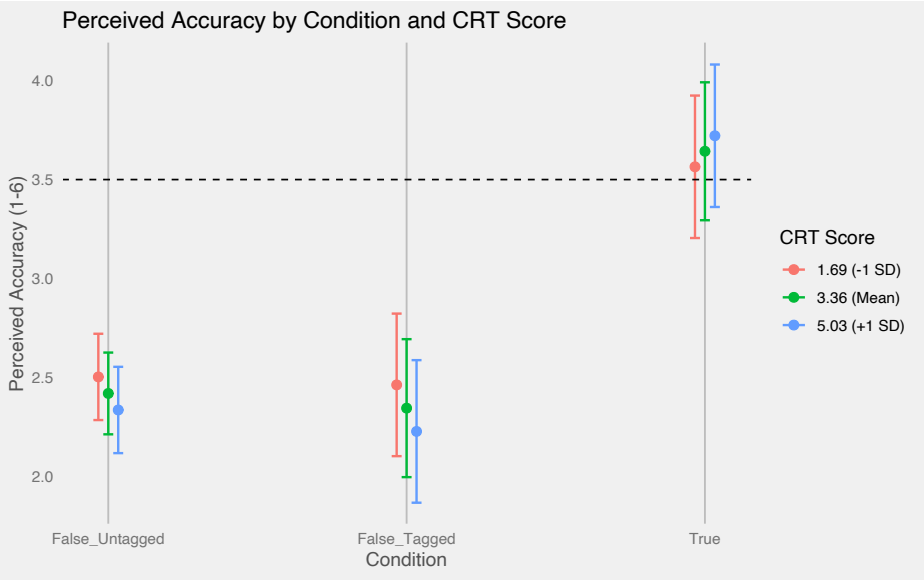

Figure 3: Predicted mean values, with 95% confidence intervals, of belief in headline by score on the cognitive reflection task. Dotted black line at 3.5 represents the mean threshold of perceived as accurate (above 3.5) or inaccurate (below 3.5)

ceive them as accurate relative to Democrats (β = 0.26, p < 0.001) and Independents (β = 0.25, p = 0.002); however, no such differences were observed for untagged false headlines between Republicans and Democrats (β = 0.09, p = 0.843) or Independents (β = 0.02, p = 0.100), suggesting the tags induced politically motivated patterns of belief. See Figure 3 for a plotting of the belief effects by party.

The same pattern of politically motivated belief was observed for the effects of self-reported ideology on belief. Ideology interacted with the main effect of tags (p = 0.002), and ideology was only associated with belief for tagged false headlines (β = 0.16, p = 0.005) such that individuals who identified as left-leaning were less likely to believe tagged false headlines than right-leaning individuals. No such linear relationship was observed for untagged false headlines (β = -0.02, p = 0.664) or true headlines (β = -0.06, p = 0.244), indicating that the presence of the tags induced politically motivated belief, similar to the pattern we observe with belief; see Figure 5 on the next page below for simple slopes. For no party or level of ideology was the contrast between belief in untagged vs. tagged false headlines significant.

3.2 Sharing

Across all partisan groups we observed a main effect of disputed tags on sharing intentions. In partial support for Hypothesis 1b, participants were significantly less likely to share tagged false headlines compared to true headlines (β = -0.24, p = 0.002) and untagged false headlines compared to true headlines (β = -0.13, p = 0.0495), but there was no difference between untagged and tagged headlines (β = 0.11, p = 0.219). However, the main effect of tags was qualified by a significant interaction of party identification (p < 0.001). All pairwise contrasts utilized Tukey-adjusted p-values. Tags reduced sharing intentions relative to true headlines for Democrats (β = -0.37, p < 0.001) and Independents (β = -0.29, p < 0.001), but not Republicans (β = -0.05, p = 0.802). When comparing untagged vs tagged false headlines, Democrats were less likely to share false tagged headlines (β = 0.20, p = 0.014), but not Independents (β = 0.15, p = 0.078) or

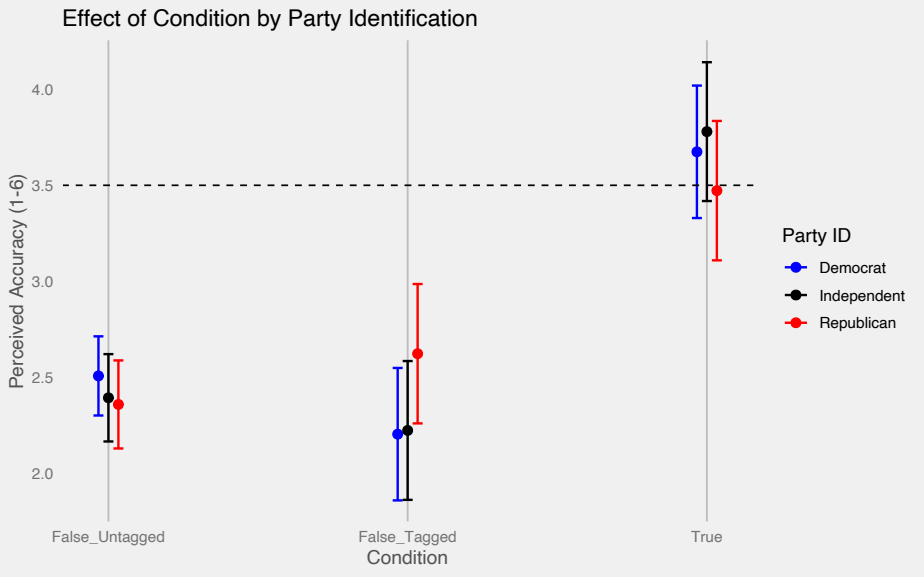

Figure 4: Predicted mean values, with 95% confidence intervals, of belief in headline by party identification. Dotted black line at 3.5 represents the mean threshold of perceived as accurate (above 3.5) or inaccurate (below 3.5)

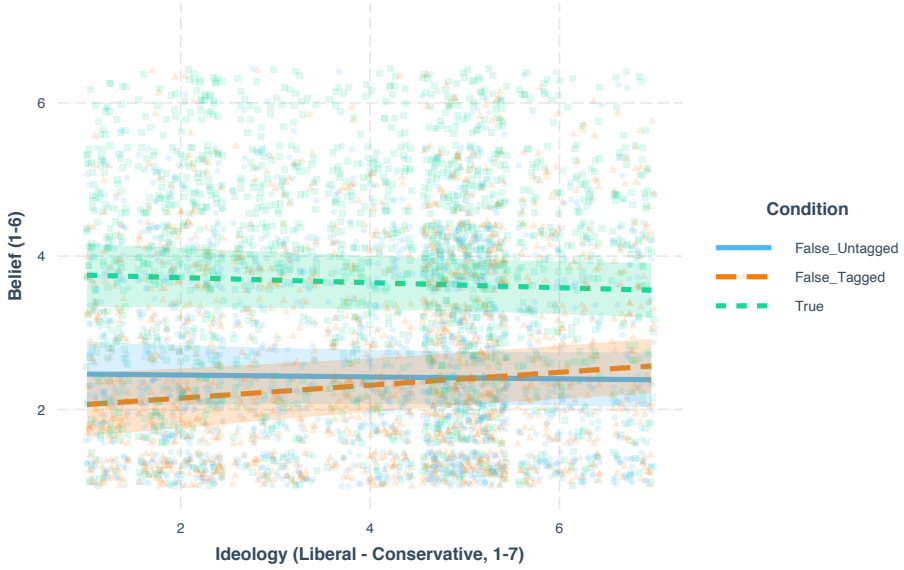

Figure 5: Simple slopes, with 95% confidence intervals, of belief in headline and ideology. Higher values on ideology measure mean more conservative, lower values more liberal. Data points are jittered.

Republicans (β = -0.03, p = 0.871). See Figure 3 on page 7 for a plotting of the sharing effects by party.

In examining the effects of cognitive reflectiveness and impulsivity, we find no support for Hypotheses 2b or 3b. There was no association of CRT performance with sharing (β = 0.07, p = 0.079), nor did CRT performance interact with condition effects (all ps > 0.270). Similarly, there was no association of BIS rating with sharing (β = 0.03, p = 0.480), nor did BIS interact with condition effects (all ps > 0.201).

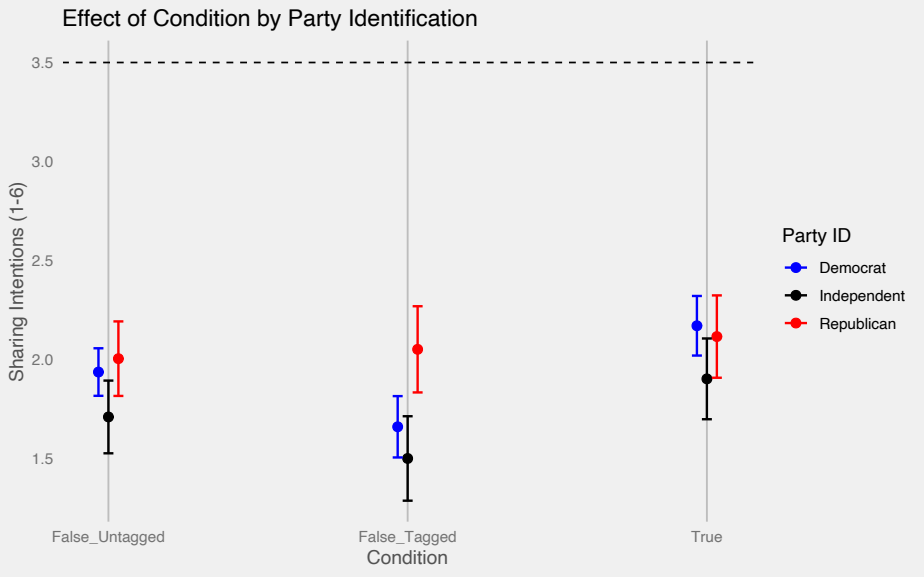

Figure 6: Predicted mean values, with 95% confidence intervals, of sharing intentions by party identification. Dotted black line at 3.5 represents the mean threshold of willingness to share a headline (above 3.5) or unwillingness (below 3.5).

To examine the role of self-reported left-right ideology, we interacted ideology with condition and found that the main effect of tags was qualified by a significant interaction (p = 0.031). Ideology was only associated with sharing intentions for tagged false headlines (β = 0.15, p = 0.014), such that individuals who identified as left-leaning were less likely to say they would share tagged false headlines than right-leaning individuals. No such relationship was observed for untagged false headlines (β = 0.04, p = 0.524) or true headlines (β = -0.01, p = 0.831), suggesting that the presence of the disputed tags caused participants to respond in a fashion aligned with their ideology—namely that liberals were less likely to express sharing intentions and conservatives more likely. We interpret this as a form of politically motivated sharing intentions. Conservatives are less trusting of social media companies than liberals (Pennycook and Rand 2019a), and we theorize that this difference in trust leads to differing evaluations of how Twitter's disputed tags may be pro-liberal/anti-conservative, leading to the observed correlation between ideology and sharing intention in the Tagged False Headline condition. See Figure 3 on page 7 for simple slopes.

In summary, while tags reduced sharing intentions relative to true (but not untagged false) among Democrats and Independents, these findings are consistent with the interpretation that disputed tags induce politically motivated intentions to share.

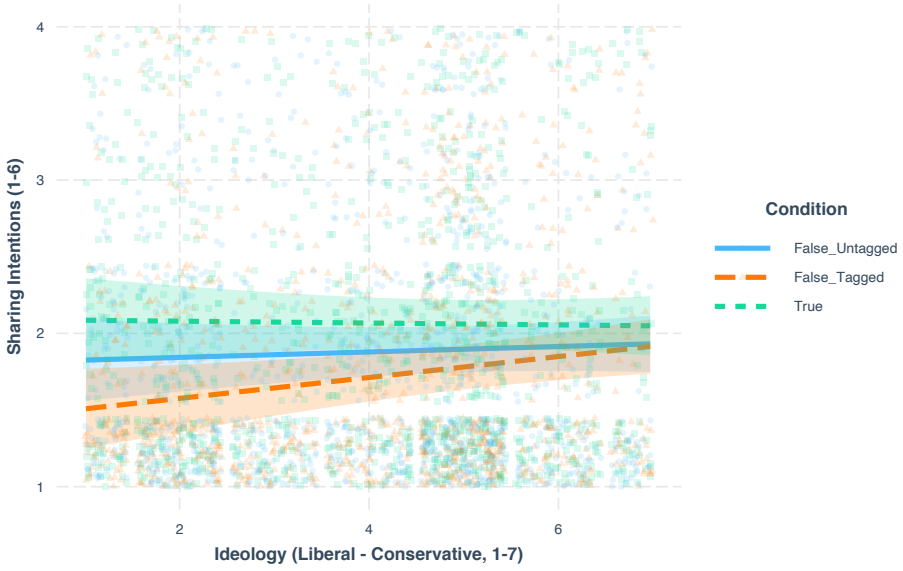

Figure 7: Simple slopes, with 95% confidence intervals, of sharing intentions and ideology. Higher values on ideology measure mean more conservative, lower values more liberal. Data points are jittered.

3.3 Headline Bias

One concern regarding our findings is that the randomized selection of headlines from the larger corpus of fake and real news headlines from Pennycook, Binnendyk, et al. (2021) may have inadvertently introduced confounding partisan biases. For example, if by chance the headlines in the False Tagged Condition were more pro-Republican, on average, than the headlines in the False Untagged Condition, that could explain observed differences in partisans' sharing intentions across the conditions, rather than said differences being the result of the tag itself. To address this potential confound, we examined the partisanship scores generated by Pennycook, Binnendyk, et al. (2021) in their testing of their corpus of headlines. Specifically, we used the "partisanship combined" score of their Democratic and Republican participants' response to the question "Assuming the above headline is entirely accurate, how favorable would it be to Democrats versus Republicans?" with the response options being "1. More favorable for Democrats, 2. Moderately more favorable for Democrats, 3. Slightly more favorable for Democrats, 4. Slightly more favorable for Republicans, 5. Moderately more favorable for Republicans, 6. More favorable for Republicans." As such, a value greater than 3.5 means the headline was viewed as pro-Republican on average, and a value less than 3.5 means the headline was viewed as pro-Democrat on average. The partisanship score for each headline used in this experiment can be found in Appendix A. We then took these scores and added them as fixed-effects to every model reported in this manuscript, to examine if our results were robust to controlling for whether the headlines were pro-Democrat or pro-Republican. All models and statistical tests regarding the partisanship of the headlines are reported in Appendix B.

In summary, our results were fully robust to controlling for headline partisanship. In no model was headline partisanship significantly associated with belief or sharing intentions. In no model did the addition of the headline partisanship predictor significantly improve model fit, and in no model did the inclusion of headline partisanship meaningfully change the other observed relationships (e.g., it never caused a previously significant coefficient to become non-significant). As such, the reported findings

in this manuscript are robust to the variance in partisanship across the text of the headlines.

4 Discussion

The present findings suggest that Twitter's disputed tags may have been relatively ineffective for many Twitter users. Although participants perceived all false headlines as less accurate than true headlines, the presence of the disputed tag did not reduce belief in false headlines.

Sharing intentions presented a slightly different story. Participants were less likely to say they would share tagged false headlines relative to true headlines, but disputed tags only reduced sharing intention relative to untagged false headlines among liberal participants.

Additionally, while performance on the cognitive reflectiveness test was related to participants' ability to identify tagged false headlines, cognitive reflectiveness was not associated with how participants responded to untagged false headlines, nor was there any evidence the tags were more effective on those low in cognitive reflectiveness vs. those high in cognitive reflectiveness. Because individuals with low cognitive reflectiveness tend to need the most help in identifying misinformation (Pennycook and Rand 2019b), ideal interventions might target those individuals. Twitter's "disputed" tags did not appear to do this. Notably, cognitive reflectiveness and self-reported impulsivity were not observably associated with sharing intentions for any headlines, false or true.

Republicans were more likely to perceive tagged false headlines as accurate relative to Democrats and Independents, and right-leaning individuals were more likely to believe the tagged false headlines compared to their left-leaning counterparts. These associations between partisanship/ideology and belief in/sharing of the headlines only arose for tagged false headlines. No such relationships were observed for true headlines or untagged false headlines. We interpret this as evidence for politically motivated sharing intentions and belief in response to the presence of a tag, and theorize that it is rooted in differing levels of trust toward social media platforms between liberals and conservatives (Pennycook and Rand 2019a). For a broader discussion of how "politically motivated" processes can be operationalized in the context of fake news research, see Pennycook and Rand (2021). Taken together, these findings suggest that the disputed tags may provide modest benefits for reducing sharing of false headlines for Democrats and Independents, but do not influence Republicans.

Overall, the disputed tags appeared to only minimally reduce sharing intentions and in contrast with other work (e.g. Pennycook, McPhetres, et al. (2020), Pennycook, Epstein, et al. (2021), and Yaqub et al. (2020)) interacted with ideology rather than reflectiveness to produce those effects. Specifically, disputed tags only reduced sharing intentions for Democrats and Independents, not Republicans. Ideology demonstrated a similar association with sharing intentions, with left-leaning participants less likely to share tagged false headlines than right-leaning participants. Similar patterns were found for perceived accuracy for the headlines. Notably, we do replicate the overall pattern of past work: participants' ability to discern true from false headlines does not translate into differences in sharing intentions.

4.1 Contributions and Limitations

A notable contribution of the present work is the use of a fully within-subject design, compared to most past work, which has used between-subjects manipulations (e.g. Clayton et al. (2020) and Pennycook, Epstein, et al. (2021)). Such past work has also used stronger stimuli (i.e., more information, visually bolder) compared to the more subtle Twitter tags. As such, the present study possessed a level of verisimilitude that past work often lacked, which may explain our divergence from past findings. For example, Clayton et al. (2020) used a between-subjects experiment to investigate the effect of Facebook's disputed tags on the perceived accuracy of fake news, and observed a moderate and significant standardized effect of 0.26. Conversely, using our within-subjects design, we observed a very small and insignificant standardized effect of 0.05, and were powered at 90% to detect standardized effects of 0.21 or greater. Most Twitter users do not view a single fact-check, then a single tweet of which they then make a judgment. Rather, they view a stream of tweets in which any potential fact-check (or other intervention) is embedded. Our results suggest that between-subject experiments where participants do not have to discern between tagged and untagged stimuli may be inflating the effect sizes of tags, and similar warnings, on belief in and willingness to share fake news. Nonetheless, our experiment was conducted in a controlled and contrived setting, so caution is needed in generalizing to how individuals respond to misinformation tags "naturally" on Twitter itself.

Despite a relatively consistent pattern of results, we caution against strong inferences to the null. While our experiment was statistically powered to reliably detect small effects by conventional definitions (i.e., d = 0.2), we were unable to detect *exceptionally* small effects, such as <d = 0.10, which have been observed in some past studies on fact-checking (e.g., Clayton et al. (2020)). Defenders of interventions with such small effects have provided compelling evidence that even when an effect is very small, if that intervention is distributed to millions of individuals at low costs then it is still beneficial (e.g., Bond et al. (2012)). That logic would certainly apply to Twitter's disputed tags (and their current "misleading" tags), and we do not rule out the possibility that the tags do have a very small beneficial effect, as our study was not powered to detect the small effect sizes observed in Bond et al. (2012). However, our results suggest that any such positive effects may occur alongside an induction of politically motivated processes. Nonetheless, our findings related to party identification and ideology were from non-preregistered analyses, which were pursued largely because many of our stated hypotheses were not supported. As such, future work should seek to replicate and extend the present findings via confirmatory tests.

Another limitation of the current work is that it may have artificially reduced base rates of sharing intentions. In choosing to use previously published stimuli, we necessarily used tweets that may have seemed old to participants, as all tweets were dated to September 2020 and the true headlines were from that month, yet participants partook in June of 2021. Also, the accuracy nudge literature finds that simply asking people to think about accuracy reduces sharing intentions (for an overview, see Pennycook and Rand (2021)), and therefore asking both perceived accuracy and sharing intentions simultaneously may have reduced sharing intentions. While these design decisions may have reduced base rates of sharing intentions, we argue that it is unlikely they would alter the relationship between sharing intentions and cognitive reflectiveness. It is also unlikely that such a potential effect would have varied across political subgroups in our sample (e.g., it absorbed the effect of the tags for Republicans only), as the literature from which our stimuli and dependent variables were drawn has consistently argued that the effectiveness of accuracy nudges is invariant to political ideology (see Pennycook and Rand (2022)). Moreover, the possibility that floor effects are concealing an

association between sharing intention and cognitive reflection is unlikely, because we do observe linear associations between ideology and sharing intentions.

Similarly, a concern with our fully within-subjects design is that the first presence of the tag effectively acts as an accuracy nudge, which then affects responses to all subsequent headlines, such that it would suppress the effect of subsequent tags relative to untagged false headlines on participant responses. Probabilistically, 92.4% of participants saw a disputed tag within the first five (of 30) headlines, and because headline order was completely randomized, robustly examining possible order effects of the first tag participants see was not plausible. While this design concern is reasonable, work on the implied truth effect would suggest the opposite effect. For example, attaching warning labels to some fake news headlines can lead to unlabeled fake news to be seen as more accurate (Pennycook, Bear, et al. 2020), which we do not observe. More broadly, we argue that our main effects are largely consistent with the accuracy nudge literature, and the ways in which our results diverge from the accuracy nudge literature do not suggest that the tags are suppressing effects one would predict based on past accuracy nudge findings. Congruent with the accuracy nudge literature, we found more evidence that our tags affect sharing intentions more than perceived accuracy (for overview, see Pennycook and Rand (2022). Unlike the accuracy nudge literature, we found that party identification and ideology were more associated with sharing intentions and perceived accuracy of tagged fake news than cognitive reflection. It is unlikely this latter finding is the result of spillover effects from participants having seen repeated tags. Rather, we interpret this finding as evidence that conservatives' decreased trust in social media platforms relative to liberals (Pennycook and Rand 2019a) led to politically motivated sharing intentions and belief in response to headlines containing the tag.

Twitter's removal of disputed tags echoes the present findings, that "this claim has been disputed" tags may have been ineffective at reducing belief in fake news. Twitter is now testing labels such as "stay informed" or "misleading" with colored icons indicating the severity of the misinformation (e.g., red exclamation point for dangerous misinformation) (Ortutay 2021). These tags may be more effective at reducing fake news sharing due to increased saliency and specific messaging (Braun, Mine, and Clayton Silver 1995). A lack of understanding was thought to be a critical issue with the disputed tags, with users confused on who "disputed" the information (Ortutay 2021). Nonetheless, it is important not to exaggerate the differences between "this tweet is misleading" and "this tweet is disputed." While it is plausible that this small change in language decreased confusion, and therefore increased the effectiveness of the tags, it is much less plausible that the "misleading" tags are interacting with different psychological mechanisms than the "disputed" tags, given their otherwise similar characteristics. If both styles of tags work through the same psychological mechanisms, the findings presented herein are likely generalizable to the "misleading" tags Twitter is now using.

The present study only explored headlines from fake news websites, yet Twitter users can also post misinformation unattached to a headline (e.g., Twitter placed "disputed" tags on several posts of then U.S. President Donald Trump). Our results may not generalize to contexts where the source of the fake news is perceived as (in)congruent with one's ideology (Traberg and Linden 2022), such as when fake news is shared by a politician (Swire-Thompson et al. 2020). Moreover, given the growing evidence that fake news is overall rare in the social media ecosystem (Allen et al. 2020; Guess, Nagler, and Tucker 2019), and that elite rhetoric may be a primary driver of belief in misinformation (Clayton et al. 2021), future research should explore how the source of misinformation may interact with disputed tags.

In the phishing context, similar warnings regarding website authenticity are more effective when the user is trained on the rationale behind the warning (Yang et al. 2017).

Thus, Twitter users may require more formal training to understand the purpose of the new tags rather than simply relying on participants to be more reflective when a tag is present. Additionally, given the partisan response to tagged headlines, it is possible that interventions will be ineffective for Republicans, who may be suspicious of any attempts to label misinformation and perceive such efforts as anti-conservative. Lastly, users may begin to ignore tags over time, especially in situations of low reliability (i.e., incorrectly tagging true information) (Parasuraman and Riley 1997). This may have contributed to the failure of the disputed tags, and presents a challenge for future interventions, regardless of their initial efficacy.

In conclusion, we find mixed evidence for the effectiveness of Twitter's disputed tags, and all potential benefits lie in their effect on Democrats and Independents. For Republicans, tags did not affect sharing intentions or belief in false headlines, and instead we found evidence that tags led liberals and conservatives to respond in a politically motivated fashion. We also found little evidence that cognitive reflectiveness or impulsivity played a meaningful role in the impact the disputed tags had on participants.

References

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. "Evaluating the fake news problem at the scale of the information ecosystem" [in en]. *Science Advances* 6, no. 14 (April): eaay3539. Accessed July 21, 2022. https://doi.org/10.1126/sciadv.aay3539. https://www.science.org/doi/10.1126/sciadv.aay3539.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. "Random effects structure for confirmatory hypothesis testing: Keep it maximal" [in en]. *Journal of Memory and Language* 68, no. 3 (April): 255–78. Accessed February 27, 2019. https://doi.org/10.1016/j.jml.2012.11.001. https://linkinghub.elsevier.com/retrieve/pii/S0749596X12001180.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization" [in en]. *Nature* 489, no. 7415 (September): 295–98. Accessed June 7, 2022. https://doi.org/10.1038/nature11421. http://www.nature.com/articles/nature11421.

Brauer, Markus, and John J. Curtin. 2018. "Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items." [in en]. *Psychological Methods* 23, no. 3 (September): 389–411. Accessed March 11, 2019. https://doi.org/10.1037/met0000159. http://doi.apa.org/getdoi.cfm?doi=10.1037/met0000159.

Braun, Curt C., Paul B. Mine, and N. Clayton Silver. 1995. "The influence of color on warning label perceptions" [in en]. *International Journal of Industrial Ergonomics* 15, no. 3 (March): 179–87. Accessed August 25, 2021. https://doi.org/10.1016/0169-8141(94)00036-3. https://linkinghub.elsevier.com/retrieve/pii/0169814194000363.

Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. "The COVID-19 social media infodemic" [in en]. *Scientific Reports* 10, no. 1 (December): 16598. Accessed March 30, 2021. https://doi.org/10.1038/s41598-020-73510-5. http://www.nature.com/articles/s41598-020-73510-5.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, et al. 2020. "Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media" [in en]. *Political Behavior* 42, no. 4 (December): 1073–95. Accessed July 13, 2021. https://doi.org/10.1007/s11109-019-09533-0. http://link.springer.com/10.1007/s11109-019-09533-0.

Clayton, Katherine, Nicholas T. Davis, Brendan Nyhan, Ethan Porter, Timothy J. Ryan, and Thomas J. Wood. 2021. "Elite rhetoric can undermine democratic norms" [in en]. *Proceedings of the National Academy of Sciences* 118, no. 23 (June): e2024125118. Accessed June 29, 2021. https://doi.org/10.1073/pnas.2024125118. http://www.pnas.org/lookup/doi/10.1073/pnas.2024125118.

Epstein, Ziv, Adam J. Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G. Rand. 2021. "Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online" [in en]. *Harvard Kennedy School Misinformation Review* (May). Accessed July 2, 2021. https://doi.org/10.37016/mr-2020-71. https://misinforeview.hks.harvard.edu/?p=7273.

Frederick, Shane. 2005. "Cognitive reflection and decision making" [in en]. *Journal of Economic Perspectives* 19, no. 4 (November): 25–42. Accessed April 1, 2021. https://doi.org/10.1257/089533005775196732. https://pubs.aeaweb.org/doi/10.1257/089533005775196732.

Garrett, R. Kelly, and Robert M. Bond. 2021. "Conservatives' susceptibility to political misperceptions" [in en]. *Science Advances* 7, no. 23 (June): eabf1234. Accessed June 6, 2021. https://doi.org/10.1126/sciadv.abf1234. https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abf1234.

Gawronski, Bertram. 2021. "Partisan bias in the identification of fake news" [in en]. *Trends in Cognitive Sciences* 25 (9): 723–24. https://doi.org/https://doi.org/10.1016/j.tics.2021.05.001.

Green, Peter, and Catriona J. MacLeod. 2016. "simr: an R package for power analysis of generalized linear mixed models by simulation" [in en]. *Methods in Ecology and Evolution* 7 (4): 493–98. Accessed August 28, 2019. https://doi.org/10.1111/2041-210X.12504. https://CRAN.R-project.org/package=simr.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook" [in en]. *Science Advances* 5, no. 1 (January): eaau4586. Accessed March 2, 2021. https://doi.org/10.1126/sciadv.aau4586. https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aau4586.

Facebook fail: Social network scraps 'disputed' flags on 'fake news'. 2017 (December). Accessed August 5, 2021. https://www.usatoday.com/story/tech/2017/12/21/facebook-fail-social-network-scraps-disputed-flags-fake-news/973153001/.

Jerit, Jennifer, and Yangzi Zhao. 2020. "Political misinformation" [in en]. *Annual Review of Political Science* 23, no. 1 (May): 77–94. Accessed July 16, 2021. https://doi.org/10.1146/annurev-polisci-050718-032814. https://www.annualreviews.org/doi/10.1146/annurev-polisci-050718-032814.

Jost, John T, Sander van der Linden, Costas Panagopoulos, and Curtis D Hardin. 2018. "Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation" [in en]. *Current Opinion in Psychology* 23 (October): 77–83. Accessed July 23, 2020. https://doi.org/10.1016/j.copsyc.2018.01.003. https://linkinghub.elsevier.com/retrieve/pii/S2352250X17302828.

Judd, Charles M., Jacob Westfall, and David A. Kenny. 2017. "Experiments with more than one random factor: Designs, analytic models, and statistical power" [in en]. *Annual Review of Psychology* 68, no. 1 (January): 601–25. Accessed March 18, 2019. https://doi.org/10.1146/annurev-psych-122414-033702. http://www.annualreviews.org/doi/10.1146/annurev-psych-122414-033702.

Kumaraguru, Ponnurangam, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2007. "Getting users to pay attention to anti-phishing education: evaluation of retention and transfer" [in en]. *Proceedings of the Anti-Phishing Working Group's 2nd Annual eCrime researchers Summit,* 70–81.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "lmerTest package: Tests in linear mixed effects models." *Journal of Statistical Software* 82 (13): 1–26. https://doi.org/10.18637/jss.v082.i13.

Lewandowsky, Stephan, and Sander van der Linden. 2021. "Countering misinformation and fake news through inoculation and prebunking" [in en]. *European Review of Social Psychology* (February): 1–38. Accessed February 24, 2021. https://doi.org/10.1080/10463283.2021.1876983. https://www.tandfonline.com/doi/full/10.1080/10463283.2021.1876983.

Maertens, Rakoen, Frederik Anseel, and Sander van der Linden. 2020. "Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects" [in en]. *Journal of Environmental Psychology* 70 (August): 101455. Accessed January 13, 2021. https://doi.org/10.1016/j.jenvp.2020.101455. https://linkinghub.elsevier.com/retrieve/pii/S0272494420303492.

Nyhan, Brendan. 2021. "Why the backfire effect does not explain the durability of political misperceptions" [in en]. *Proceedings of the National Academy of Sciences* 118, no. 15 (April): e1912440117. Accessed May 31, 2022. https://doi.org/10.1073/pnas.1912440117. https://pnas.org/doi/full/10.1073/pnas.1912440117.

What's in a tag? Twitter revamps misinformation labels. 2021 (July). Accessed August 5, 2021. https://apnews.com/article/health-coronavirus-pandemic-election-2020-misinformation-technology-37cee761758f1072f91e1a362f769f5e.

Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A subject pool for online experiments" [in en]. *Journal of Behavioral and Experimental Finance* 17 (March): 22–27. Accessed November 10, 2020. https://doi.org/10.1016/j.jbef.2017.12.004. https://linkinghub.elsevier.com/retrieve/pii/S2214635017300989.

Parasuraman, Raja, and Victor Riley. 1997. "Humans and automation: Use, misuse, disuse, abuse." Publisher: SAGE Publications Sage CA: Los Angeles, CA, *Human factors* 39 (2): 230–53.

Patton, Jim H., Matthew S. Stanford, and Ernest S. Barratt. 1995. "Factor structure of the barratt impulsiveness scale." *Journal of Clinical Psychology* 51 (6): 768–74.

Pennycook, Gordon, Adam Bear, Evan T. Collins, and David G. Rand. 2020. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings" [in en]. *Management Science* 66, no. 11 (November): 4944–57. Accessed September 13, 2021. https://doi.org/10.1287/mnsc.2019.3478. http://pubsonline.informs.org/doi/10.1287/mnsc.2019.3478.

Pennycook, Gordon, Jabin Binnendyk, Christie Newton, and David G. Rand. 2021. "A practical guide to doing behavioral research on fake news and misinformation" [in en]. *Collabra: Psychology* 7, no. 1 (July): 25293. Accessed August 21, 2021. https://doi.org/10.1525/collabra.25293. https://online.ucpress.edu/collabra/article/7/1/25293/117809/A-Practical-Guide-to-Doing-Behavioral-Research-on.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. "Shifting attention to accuracy can reduce misinformation online" [in en]. *Nature* (March). Accessed March 22, 2021. https://doi.org/10.1038/s41586-021-03344-2. http://www.nature.com/articles/s41586-021-03344-2.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G. Rand. 2020. "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention" [in en]. *Psychological Science* 31 (7): 770–80.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G. Lu, and David G. Rand. 2021. *sj-docx-3-pss-10.1177_0956797620939054 – Supplemental material for Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention.* Letter to the Editor, December. https://doi.org/10.25384/SAGE.12594110.v2. https://sage.figshare.com/articles/journal_contribution/Pennycook_Supplemental_Material_rev_Supplemental_material_for_Fighting_COVID-19_Misinformation_on_Social_Media_Experimental_Evidence_for_a_Scalable_Accuracy-Nudge_Intervention/12594110.

Pennycook, Gordon, and David G. Rand. 2019a. "Fighting misinformation on social media using crowdsourced judgments of news source quality" [in en]. *Proceedings of the National Academy of Sciences* 116, no. 7 (February): 2521–26. Accessed June 6, 2022. https://doi.org/10.1073/pnas.1806781116. https://pnas.org/doi/full/10.1073/pnas.1806781116.

———. 2019b. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning" [in en]. *Cognition* 188 (July): 39–50. Accessed September 22, 2020. https://doi.org/10.1016/j.cognition.2018.06.011. https://linkinghub.elsevier.com/retrieve/pii/S001002771830163X.

———. 2021. "The psychology of fake news" [in en]. *Trends in Cognitive Sciences,* 15. https://doi.org/https://doi.org/10.1016/j.tics.2021.02.007.

———. 2022. "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation" [in en]. *Nature Communications* 13 (2333). Accessed May 3, 2022. https://doi.org/10.31234/osf.io/v8ruj. https://osf.io/v8ruj.

Pereira, Andrea, Elizabeth Harris, and Jay J Van Bavel. 2021. "Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news" [in en]. *Group Processes,* 24.

Romer, Daniel, and Kathleen Hall Jamieson. 2020. "Conspiracy theories as barriers to controlling the spread of COVID-19 in the U.S." [in en]. *Social Science & Medicine* 263 (October): 113356. Accessed March 30, 2021. https://doi.org/10.1016/j.socscimed.2020.113356. https://linkinghub.elsevier.com/retrieve/pii/S027795362030575X.

Roth, Yoel, and Nick Pickles. 2020. "Updating our approach to misleading information." *Twitter Blog* (May). Accessed July 29, 2022. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

Smith, Jeff. 2017. "Designing against misinformation." *Design at Meta* (December). Accessed July 29, 2022. https://medium.com/designatmeta/designing-against-misinformation-e5846b3aa1e2.

Swire-Thompson, Briony, Joseph DeGutis, and David Lazer. 2020. "Searching for the backfire effect: Measurement and design considerations" [in en]. *Journal of Applied Research in Memory and Cognition* 9, no. 3 (September): 286–99. Accessed July 24, 2021. https://doi.org/10.1016/j.jarmac.2020.06.006. https://linkinghub.elsevier.com/retrieve/pii/S2211368120300516.

Swire-Thompson, Briony, Ullrich K. H. Ecker, Stephan Lewandowsky, and Adam J. Berinsky. 2020. "They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation" [in en]. *Political Psychology* 41, no. 1 (February): 21–34. Accessed March 2, 2021. https://doi.org/10.1111/pops.12586. https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12586.

Thomson, Keela S, and Daniel M Oppenheimer. 2016. "Investigating an alternate form of the cognitive reflection test" [in en]. *Judgment and Decision Making* 11 (1): 15.

Traberg, Cecilie Steenbuch, and Sander van der Linden. 2022. "Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility" [in en]. *Personality and Individual Differences* 185 (February): 111269. Accessed February 17, 2022. https://doi.org/10.1016/j.paid.2021.111269. https://linkinghub.elsevier.com/retrieve/pii/S0191886921006486.

Twitter Support, @TwitterSupport. 2021. *Last year, we started using labels to let you know when a Tweet may include misleading information. For some of you on web, we'll be testing a new label design with more context to help you better understand why a Tweet may be misleading [Tweet].,* July. Accessed July 29, 2022. https://twitter.com/TwitterSupport/status/1410646566885601280.

Walter, Nathan, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. "Fact-checking: A meta-analysis of what works and for whom" [in en]. *Political Communication* 37, no. 3 (May): 350–75. Accessed September 2, 2020. https://doi.org/10.1080/10584609.2019.1668894. https://www.tandfonline.com/doi/full/10.1080/10584609.2019.1668894.

Wood, Thomas, and Ethan Porter. 2019. "The elusive backfire effect: Mass attitudes' steadfast factual adherence" [in en]. *Political Behavior* 41, no. 1 (March): 135–63. Accessed September 17, 2020. https://doi.org/10.1007/s11109-018-9443-y. http://link.springer.com/10.1007/s11109-018-9443-y.

Yang, Weining, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. 2017. "Use of phishing training to improve security warning compliance: Evidence from a field experiment" [in en]. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp on - HoTSoS,* 52–61. Hanover, MD, USA: ACM Press. Accessed August 25, 2021. https://doi.org/10.1145/3055305.3055310. http://dl.acm.org/citation.cfm?doid=3055305.3055310.

Yaqub, Waheeb, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. "Effects of credibility indicators on social media news sharing intent" [in en]. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–14. Honolulu HI USA: ACM, April. Accessed September 13, 2021. https://doi.org/10.1145/3313831.3376213. https://dl.acm.org/doi/10.1145/3313831.3376213.

Authors

Jeffrey Lees was a Visiting Assistant Professor in the Department of Economics at Clemson University at the time this research was conducted, and is currently an Associate Research Scholar at Princeton University's Andlinger Center for Energy and the Environment.

Abigail McCarter was an undergraduate student at Clemson University at the time this research was conducted, and is currently the Development and Impact Operations Manager at the Truman Center for National Policy and Truman National Security Project, and a master's candidate at Johns Hopkins School of Advanced International Studies.

Dawn Sarno is an Assistant Professor of Psychology at Clemson University.

Acknowledgements

We thank all participants in the Spring 2021 Clemson University course Phishing for Trolls, Bots, and Hackers: The Generalized Analysis of Online Inauthenticity for their feedback on study design, Dave Rand for feedback on early versions of the manuscript, and attendees to the Watt Family Center's April Fools Symposium at Clemson University and the Symposium on Uncommon Yet Consequential Online Harms at the Stanford University's Internet Observatory for their helpful comments and questions on our findings.

Data Availability Statement

All data, analysis code, and study materials are publicly available on the Open Science Framework at https://osf.io/h65nv/. Jeffrey Lees prepared all the open materials, code, and data; do not hesitate to reach out to them with any questions.

Funding Statement

This project was supported by the Clemson University Creative Inquiry program, through the Clemson University Media Forensics Hub. All funds were used to compensate participants.

Competing Interests

The authors declare no competing interests.

Ethical Standards

This research was approved by Clemson University's Institutional Review Board. All participants provided informed consent to participate. No deception was utilized in this research, and all participants were debriefed at the end of the study where the false nature of the headlines was made explicitly clear.

Keywords

misinformation; Twitter; partisanship; fake news; cognitive reflection.