“Fact-checking” videos improve truth discernment ability but do not reduce fake news sharing on Twitter

Alexander Bor∗1, Mathias Osmundsen1, Stig Hebbelstrup Rye Rasmussen1, Anja Bechmann2, and Michael Bang Petersen1

1Department of Political Science, Aarhus University
2School of Communication and Culture, Aarhus University

September 22, 2020

Abstract
“Fake news” are widely acknowledged as an important challenge for Western democracies. Yet, surprisingly little effort has been devoted to measuring the effects of various counter-strategies. We address this void by running a pre-registered field experiment analyzing the causal effects of popular fact-checking videos on both believing and sharing fake news among Twitter users (N = 1,600). We find that the videos improve truth discernment ability as measured by performance in a fake news quiz immediately after exposure. However, the videos have not reduced sharing links from verified “fake news” websites on Twitter in the weeks following the exposure. Indeed, we find no relationship between truth discernment ability and fake news sharing. These results imply that the development of effective interventions should be based on a nuanced view of the distinct psychological motivations of sharing and believing “fake news”.

Significance statement
“Fake news” spread on social media because ordinary citizens are eager to share them. Hence, educational interventions against “fake news” seek to empower citizens to identify and abstain from sharing such information. These interventions assume that the psychological factors underlying the well-studied phenomenon of belief in “fake news” also shape sharing decisions. Here, we demonstrate that educational interventions reduce beliefs in but not the sharing of “fake news” on Twitter. We also show that the people who fall for “fake news” are not those who share them. These findings may help to design more efficient tools against “fake news” as we face unprecedented waves of misinformation in the context of the COVID-19 pandemic and electoral turmoil in the US.

Word count: 5,644

This working paper has not been peer-reviewed yet.

∗Corresponding author: alexander.bor@ps.au.dk
The circulation of false and misleading news stories (often referred to as “fake news”) is perceived as a major problem in Western democratic politics. Misinformation may have played a central role in recent political events such as the 2016 Presidential election in United States and the Brexit referendum in Great Britain (Allcott and Gentzkow 2017; Marshall and Drieschova 2018). Consistent with this, research finds that people are inclined to believe misinformation that fits their political worldview (Miller, Saunders and Farhart 2016). While “fake news” is often created by elite actors for financial or political gain and often relies on bot networks for initial promotion, lay people are instrumental in facilitating their spread. As false news stories are often both novel and emotionally engaging, social media users share them “faster and further” than accurate news stories (Vosoughi, Roy and Aral 2018). While the overall prevalence of false news stories is low (Guess, Nagler and Tucker 2019), research shows that single stories can attract significant public attention and that false information may be particular likely to receive circulation in the situations where it may be most dangerous: at times of high social polarization (Horowitz 2001).

Recent years have witnessed an explosion in collaborative efforts to stem the tide of “fake news.” Computer scientists and tech companies like Facebook and Twitter are exploring ways to remove bots and “sock puppets” from their user base and to adjust their page-ranking algorithms to filter out fabricated news stories (Crowell 2017; Weedon, Nuland and Stamos 2017). Data scientists are developing machine learning techniques that rely on text classification and network analysis to detect and stop “fake news” stories before they go viral on social media (Conroy, Rubin and Chen 2015; Shu, Bernard and Liu 2019; Tacchini et al. 2017), while educators and journalists are creating training programs and other resources that teach schoolchildren to evaluate information credibility (Jones 2017). The European Union has invested millions of euros in projects like the Social Observatory for Disinformation and Social Media Analysis (SOMA) and the European Digital Media Observatory (EDMO) that evaluate counter-strategies against the spread and impact of online fake news, and the East Stratcom Task Force, helping civil society build resilience against Russian disinformation campaigns. Perhaps most prominently, media companies across the world are investing heavily in fact-checking programs (Graves 2016) that, according to communication research, can hinder the spread of “fake news” (Nyhan and Reifler 2015).
The promise of these advances notwithstanding, the proliferation of “fake news” poses a major challenge. Tech companies and scientists are in an arms race against “fake news” producers who use increasingly ingenious and sophisticated tools to disseminate misinformation. And constraints on time and resources put pressure on professional fact-checkers to keep up with the latest “fake news” stories before they hit social media. For these reasons, social scientists have increasingly turned to prebunking rather than debunking strategies (van der Linden 2019; Guess et al. 2020). In other words, they seek to equip citizens with broadly applicable tools to recognize “fake news” stories in order to prevent belief and sharing, instead of correcting beliefs after the damage is already done.

Media organizations and researchers have converged on a particular strategy to empower citizens: Minutes-long entertaining educational programs (Moyer-Gusé 2008) that effectively turn citizens into fact-checkers by teaching the core principles of spotting “fake news”. Simple and effective tools include reading beyond the headline, googling the headline or main claims, and verifying the credibility of the source and the author. Multiple stakeholders, like governmental and non-governmental institutions, and major news sites have created their own versions of such videos, and they have been watched by millions (see below). Educational videos and other teaching materials are also at the heart of the EU’s strategy to raise awareness about “fake news” in civil society and to enable citizens to navigate the online news environment safely (EC 2018). Yet, despite high hopes for the role they can play in fighting “fake news”, the educational interventions lack rigorous program evaluation: We simply do not know how effective they are. In the present manuscript, we provide the first behavioral-experimental test of the effectiveness of these educational interventions in decreasing people’s propensity to share “fake news” on a social media platform, Twitter. Before turning to the results of our program evaluation, however, we briefly discuss the main theoretical arguments for and against the effectiveness of educational interventions.

Empowering citizens to counter the threat of “fake news” necessitates an understanding of the psychological factors shaping the decision to share false news stories. Convincing someone to stop sharing “fake news” would require different tools depending on whether they believe these news and share them with good faith, whether they find them funny and post them to entertain their (potentially unassuming) followers or whether they share to smear dirt on their
partisan opponents (Osmundsen et al. 2020). Educational fact-checking interventions assume (often implicitly) that people are motivated by a particular set of motivations, which social psychologists call *accuracy motivations* (Pennycook et al. 2020). Accuracy motivations imply that people are motivated to hold accurate beliefs about the world (Chaiken and Trope 1999). In the context of social media, these accuracy motivations imply that the average user wants to share accurate news but may not be motivated or competent enough to assess the veracity of news (Cappella, Kim and Albarracín 2015; Pennycook et al. 2020). From this perspective, educational interventions work because they make citizens better able to determine whether news stories are true or false. In line with these arguments, we know that individuals who invest more cognitive resources in problem-solving (Pennycook and Rand 2019b) and who enjoy analytical thinking (Pennycook and Rand 2018) are less likely to fall prey to “fake news.” Elderly social media users – arguably due to shortfalls in digital media literacy and a failure to discern “fake” from “real” online – are among the worst offenders when it comes to sharing “fake news” on multiple social media platforms (Guess, Nagler and Tucker 2019; Osmundsen et al. 2020). Finally, Guess et al. (2020) recently demonstrated that exposure to Facebook’s “Tips to Spot False News,” a short list of simple tools for identifying misleading news, improved the ability to discern between false and mainstream news story headlines; an improvement that remained measurable more than two weeks later after the intervention.

Yet, to the best of our knowledge, no prior study offers direct evidence that the inability to discern “fake” from real news contributes to “fake news” sharing. Hence, the key psychological premise underlying fact-checking and related interventions remains untested. This is especially problematic as illiteracy is not the only potential psychological mechanism behind “fake news” sharing. The literature on group psychology suggests that when people process politically relevant information, they are less concerned about the veracity of news and more about their usefulness for derogating disliked groups (Mercier 2020). For example, people tend to believe more in conspiracy theories if they fit their ideological or partisan dispositions, and these effects are larger among those most knowledgeable about politics (Miller, Saunders and Farhart 2016). Furthermore, people are motivated to share “fake news” that portrays disliked groups negatively, in some cases even if they do not believe the news is true (Petersen, Osmundsen and Arceneaux 2018). If people are undisturbed by the mendacity of the news they share online, efforts to
increase their ability or motivation to establish veracity may be ineffective. In fact, Guess et al. (2020) show that these concerns are valid: While their media literacy intervention strengthened the ability to detect false news, additional analyses suggested that it failed to lower self-reported intentions to share “fake news.”

On these grounds, our aim is to conduct a direct behavioral test of whether prominent educational fact-checking videos reduce “fake news” sharing on Twitter by improving the ability to discern between information and misinformation. In doing so, we assess the psychological model underlying their presumed effectiveness. Specifically, we subject the following questions to empirical test: (1) Does exposure to educational fact-checking interventions increase the ability to assess the veracity of news stories (i.e. truth discernment)? (2) Does exposure to educational fact-checking interventions decrease actual sharing of “fake news” stories on Twitter? (3) Is the ability to assess the veracity of news stories related to actual sharing of false news stories on Twitter?

To answer these questions, we first ran a pilot study on Amazon’s Mechanical Turk (N = 686). Next, to obtain results from a highly naturalistic and ecologically valid environment, we ran a pre-registered field experiment on a unique panel of Twitter users (N = 1,600). Respondents in the treatment arm were randomly assigned to watch one of six professionally produced fact-checking videos from government-funded or non-governmental organizations (e.g., “How to Spot Fake News – Factcheck.org”, see Table 1). Thus, the utilized experimental materials have maximal ecological validity, and we are testing the effectiveness of the materials actually available to social media users at scale. These short (<10 min) fact-checking videos highlighted widely acknowledged anti-misinformation practices like the importance of considering source credibility and searching the Internet for supporting evidence. We examined in both the MTurk and the Twitter samples whether respondents who watched the videos were better at assessing the veracity of news headlines in a fake news quiz than participants who did not watch the videos.\footnote{The quiz items are developed by Pennycook et al. (2020). We ran a pilot study on Lucid (N = 997) and employed item-response theory models to select the eight best items from a battery of 75. See more details in Section C.} Relying on the Twitter sample, we examined the real-world implications of the intervention by testing whether the educational videos made respondents less likely to share links to “fake news” sources on Twitter. In testing this claim, we followed best practices (Guess, Nagler and Tucker 2019; Grinberg et al. 2019; Lazer et al. 2018) and identified “fake news” on
the source level. In other words, we operationalized “fake news sharing” as (re)tweeting URLs from websites with a known history of publishing factually incorrect news (for a list of URLs, see [Alcott and Gentzkow 2017] Guess, Nyhan and Reifler 2018).

Results

Educational videos improve truth discernment

First, we find that the educational videos work as intended; they increase the ability to differentiate between fake and real news. Compared to participants in the control group, Twitter: $M(SD) = 0.37(0.24)$; MTurk: $M(SD) = 0.30(0.23)$, participants who watched the educational videos, Twitter: $M(SD) = 0.42(0.25)$; MTurk: $M(SD) = 0.38(0.25)$ had significantly higher truth discernment scores, Twitter: $\Delta M = 0.05$, 95%CI = (0.02; 0.08), $p < 0.01$, Cohen’s $d = 0.2$; MTurk: $\Delta M = 0.07$, 95%CI = (0.02; 0.13), $p < 0.05$, Cohen’s $d = 0.3$. Although these effects may appear modest, it is worth noting that the average difference between participants with and without higher education in our Twitter sample is not much larger ($\Delta M = 0.08$, 95%CI = (0.06; 0.11), $t(991) = 6.8$, $p < 0.001$, Cohen’s $d = 0.36$).

Figure 1: Educational videos improve truth discernment ability both among Twitter and MTurk users

![Graph showing educational videos improve truth discernment ability](image)

Note: The figure shows coefficient estimates from multilevel models with random slopes for videos along with 90% and 95% confidence intervals.
Figure 1 plots these pooled estimates along with estimates for each of our six videos. It demonstrates that there is little systematic variation across the videos. While all of them have similar, statistically significant effects (with a single exception), some of them had above average effects in the MTurk sample, others in the Twitter sample. John Spencer and MediaSmarts produced the two videos with slightly above average effects in both samples.

As a post-hoc analysis, we disaggregate discernment scores into a measure of “gullibility” – the average perceived accuracy of fake news – and “trust in mainstream media” – the average perceived accuracy of real news. Regressing our treatments on these measures shows that the educational materials have the intended effects on both attributes, although the decrease in gullibility is somewhat larger (around 0.04 points) than the increase in mainstream media trust (between 0.01 and 0.03 points).

**Educational videos do not reduce fake news sharing**

Next, we turn to the effect of educational videos on fake news sharing on Twitter. Modeling fake news sharing is a challenge as it is a rare, erratic behavior. Our respondents shared 2,319 “fake news” in the focal fourteen months of our study, and only 12% of participants (189 users) shared any fake news in the study period. We employ three important tools to improve the accuracy of our estimates in our data: 1) we employ a difference-in-differences design, which allows us to ask whether the treatment makes a difference taking into account differences between the treatment and control groups in the pre-treatment period and the behavior change from pre- to post-treatment in the control group; 2) we rely on three different conceptualizations of “fake news” sharing behavior, each mitigating different aspects of methodological issues; 3) finally, we aggregate our data to user-weeks, instead of more granular units (e.g. user-days), which reduces the variance of our data substantially. See Figure OAI for time trends in our outcome variables.

Our main model contrasts the number of fake news shared in the three months prior to and nine weeks after the intervention. We find that the treatment group shared 55% more “fake news” than expected, although the estimate is quite noisy ($\beta = 1.55; 95\% CI : [0.93 - 2.58]$).

\footnote{In our pre-registration, we planned to observe three months of post-treatment behavior on Twitter, but due to a programming error we had to limit our inquiry to nine weeks. Analyzing the third month of behavior after exposure could have been interesting to observe decay. But as we see no short-term effects, this is less of a concern.}
Still, given this model and our data, a reduction beyond 7% in number of fake news shared with high confidence is unlikely.

**Figure 2**: Treatment effect on sharing “fake news” by various pre-treatment periods (vertical panels) and post-treatment periods (x axis). All coefficients are odds ratios based on the interaction terms in difference-in-differences models.

Note: Number of fake news modeled in a negative binomial model, Probability of sharing one or more fake news and Proportion of fake news are modeled as binomial models. Error bars denote 90% and 95% credible intervals.

To test the substantive and methodological robustness of our findings, we report several additional models in Figure 2. First, we explore both shorter and longer pre-treatment periods from one year to two weeks (depicted as vertical facets). Second, we explore shorter post-treatment periods from one to nine weeks (depicted on x axis). Third, we test the effect for two additional outcome measures: 1) the probability that the respondent shares at least one fake news story in a given week, and 2) the proportion of fake news to real news shared on a given week.

Three main findings stand out. Most importantly, none of our models offer firm evidence...
that our intervention reduced fake news sharing or improved the quality of news shared by our respondents. Our results also demonstrate that longer pre-treatment periods reduce the uncertainty in our models. Finally, and unsurprisingly, our models are more confident when we explore longer post-treatment periods. Importantly, we believe it is both theoretically and normatively justified to explore long-term effects. While it may strike the reader as naive to expect an effect over two months after a five-minute intervention, it is important to remember that the explicit goal of the utilized videos is to increase respondents’ literacy. If the videos were effective in teaching viewers how to avoid posting fake news, they should have large and long-term effects. Indeed, previous research has found that effects of a digital literacy intervention on truth discernment persisted for at least three weeks (Guess et al. 2020).

What are the effects we can confidently rule out based on these models? Focusing on models with at least three months of pre-treatment and at least five weeks of post-treatment, we find that the lower boundary of the 95% credible intervals are between a 36% decrease and a 4% increase in number of fake news items shared, and between a 38% decrease and a 5% increase in the probability of sharing at least one fake news item. Interestingly, when it comes to the proportion of fake to real news, our more informative models suggest that a decrease larger than 29% is unlikely. Notably, some of our models here find a statistically significant increase in the proportion of fake news.

Admittedly, because of the noisiness of the data, we cannot rule out that the intervention had a meaningful – albeit small – positive effect decreasing “fake news” sharing. As a post-hoc robustness check, we also estimate the average quality of news shared for every respondent in every week, relying on expert ratings from Pennycook and Rand (2019a). This yields a more nuanced, continuous measure of news shared taking into account “real”, “fake” and hyper-partisan sources. At the same time, this analysis limits the data we can take into account. First, we succeeded in matching 49 of the 60 news sources with expert ratings. These sources capture 42% of our news links. Second, many respondents do not share any political news and are excluded from the analysis. This leaves us with 590 respondents for this analysis.

Figure 3 shows the estimates from similar difference-in-differences models as above. These results are consistent with our previous findings but are much more precise. They indicate that even the week after the intervention, the average quality of the news posted was about
Figure 3: Treatment effect on average news source quality by various pre-treatment periods (vertical panels) and post-treatment periods (x axis). All coefficients are based on the interaction terms in a difference-in-differences model.

Note: Error bars denote 90% and 95% credible intervals.

the same as before the intervention (taking into account the pre-treatment differences between treatment and control, as well as the trend in the control between pre- and post-treatment period). If we look at the upper boundaries of the estimates, anything beyond a .13 points (or half a standard deviation) increase in average news quality is unlikely. The models considering longer post-treatment periods are inconsistent with an improvement of just .05 points (or one fifth of a standard deviation). To sum up, we fail to find any evidence that the “fact-checking” videos have a positive effect on news sharing behavior, and we can rule out medium to small effects depending on the model.

Sharing is not believing

How come our intervention succeeded in improving truth discernment but failed to reduce fake news sharing on Twitter? Skeptical readers may suspect that our null finding is a false negative or a methodological artefact. Yet, we consider it notable that to the best of our knowledge, it has never been tested directly whether truth discernment is related to “fake
news” sharing behavior on social media. If this relationship turns out to be weak or nonexistent, educational “fact-checking” videos offer a remedy based on a faulty diagnosis. This is exactly what our results suggest.

We regress the number of “fake news” shared on truth discernment scores across three negative binomial regression models, 1) without any covariates, 2) only demographic covariates, and 3) with both demographic and psychological covariates (see Section D.3 in the OA). None of these models find convincing evidence for a substantial relationship between truth discernment and “fake news” sharing. While the model without covariates finds that respondents with a perfect quiz performance post on average less than half the “fake news” compared to respondents who responded randomly to all quiz questions, the model cannot rule out that they post half as many. Moreover, by adding our demographic and psychological covariates, the asymmetries in fake news sharing between the best and worst scores in the fake news quiz disappear. Our models find even less evidence for a systematic relationship between truth discernment and the likelihood of posting at least one “fake news” story during the study period. Across all three models, the coefficients estimates are close to zero.

As expected, we find a strong positive relationship between truth discernment and posting real news. Respondents with perfect truth discernment scores post seven to 16 times more real news than those who are not better than chance at truth discernment. Importantly, this relationship need not be causal: an interest in politics could lead both to better performance in the fake news quiz and to more frequent posting of news. Accordingly, political knowledge is the strongest psychological correlate of the number of real news items shared on Twitter. This could explain why we find no positive effect of our interventions on real news sharing.

Finally, the present data provides a unique opportunity to highlight notable asymmetries in the correlates of believing and sharing fake news among Twitter users. Figure 4 offers simple descriptive comparisons between factors known to shape truth discernment (digital literacy and cognitive reflection) and factors known to shape fake news sharing (age and partisan identity). On panel A1, we show that Hargittai and Hsieh’s self-reported digital literacy scale is strongly associated with truth discernment. People most familiar with digital technologies have substantially higher truth discernment scores than those with lowest digital literacy. We also

---

3As this is a post-hoc analysis, and we analyze the psychological antecedents of “fake news” sharing with partially overlapping data reported elsewhere, we limit our analysis to simple descriptive comparisons here.
**Figure 4:** Asymmetries in the correlates of believing (column 1) and sharing (column 2) “fake news”

Note: The lines are simple linear (cog.refl., partisan identity) or loess (dig.lit., age) trendlines with 95% confidence intervals.
replicate the finding that cognitive reflection (as measured by the CRT2 scale, displayed in panel B1) correlates with truth discernment (Pennycook and Rand 2019). Importantly, however, there appears to be no asymmetry between people with high versus low digital literacy or cognitive reflection when it comes to fake news sharing (panels A2 and B2).

*Vice versa*, when it comes to documented antecedents of fake news sharing, such as age (panel C2) and partisan identity (panels D2 and E2), we find no corresponding asymmetries in truth discernment. Remarkably, our oldest participants exhibit just as much skill in discerning between fake and real news as our youngest participants (panel C1) (see also Brashier and Schacter 2020). Similarly, partisanship is a strong correlate of sharing partisan “fake news” both among Democrats (D2) and Republicans (E2). Yet, contrary to the logic of partisan motivation, Democrats actually perform better at discerning between pro-Democratic “fake” and real news than Republicans (D1). Similarly, partisan motivations do not undermine Republicans’ ability to tell true from false when it comes to news that is congruent with their party’s agenda (E1).

Overall, these asymmetries highlight the important – yet underappreciated – fact that those who believe and those who share “fake news” are likely to be very different people.

**Discussion**

Stemming the tide of fake news propagated on social media is a daunting task. Solving it requires extensive, interdisciplinary efforts. Here, we have taken the first stab at evaluating one counter-strategy that media organizations and researchers have enthusiastically embraced: Educational fake news videos aimed at empowering citizens to become their own fact-checkers.

As anticipated by past work arguing that the failure to think analytically drives beliefs in fake news (Pennycook and Rand 2018), we found that the educational materials on how to spot fake news improved participants’ ability to assess the credibility of news story headlines. These findings suggested that the fact-checking videos helped participants build immediate resistance against misinformation. This is an important and, arguably, desirable finding: Fact-checking videos may serve as an effective intervention to reduce belief in “fake news.”

Yet, while these widely available educational interventions may have succeeded in lowering beliefs in “fake news,” they did not prevent participants from sharing false and untrustworthy news sources on Twitter. Worse still, participants performing well on the “fake news” quiz were
just as likely to share untrustworthy news stories. These results challenge the effectiveness of educational interventions in decreasing the circulation of “fake news”. The videos make it easier for citizens to judge the quality of information, but that does not, as is often assumed, translate into a lower likelihood of sharing false information. As social media users’ decisions to share information with friends and followers play an important role for the circulation of “fake news,” this finding is cause for concern.

At this stage, there are multiple plausible explanations for our findings. A first explanation concerns the issue of power. Fake news is a such rare phenomenon that even a large online panel of 1600 individuals has limited power to detect treatment effects. We have employed multiple design and statistical tools to alleviate these concerns, but it remains true that we can rule out smaller effects only if we look at multiple months of behavior after the treatment. Future research should seek to further improve the precision of treatment effects. Increasing sample size is an obvious but expensive solution. An alternative approach could be recruiting respondents with a history of posting fake news (see Munger [2017], Siegel and Badaan [2020]).

Another possibility concerns problems with our chosen interventions. While we intentionally selected the interventions that are currently available at scale, these may be too brief to trigger long-term changes in resistance to misinformation. While we sympathize with this argument, it places practitioners in a difficult dilemma. From an applied point of view, short, entertaining educational materials are the types of sustainable interventions that can be cost-effectively produced and distributed widely. Is it reasonable to expect citizens at large to accept and follow more extensive measures, like multi-hour educational courses? Gamified online education tools (Roozenbeek and van der Linden [2019]) have shown some promise, but their effectiveness in an externally valid behavioral setting is yet to be tested.

Finally, an entirely different explanation argues that the prevalent theoretical assumptions behind “fake news” sharing need to be revisited. Recent research warns that citizens have other goals in mind than spreading accurate and credible information when sharing news stories online. Politically motivated reasoning propels people to share politically congenial news that will help them protect their political identity. Posting outrageous stories about partisan opponents could be seen as a valuable service to fellow partisans, or a tool to signal loyalty by burning bridges to competing coalitions (Mercier [2020]). These competing theories offer a good explanation for
why the interventions failed: partisan motivations create benefits to posting misinformation, which may outweigh the potential costs of spreading false rumors.

The distinction between accuracy and partisan motivations reconciles the seemingly contradictory finding that the educational videos improved performance on the fake news quiz but failed to alter real-world sharing behavior. In an anonymous survey quiz where participants are told to carefully scrutinize news story headlines, accuracy motivations may loom large. But in social media settings – where friends and followers pay close attention to what you tweet and where social status is at stake – incentives may shift towards partisan goals. That context determines when different motivations switch on and off drives home the point highlighted earlier: Program evaluations of fake news interventions need to take place in realistic and naturalistic social media contexts. Artificial “lab” contexts may lead to very different and biased conclusions on how interventions operate in real life.

Acknowledging the importance of identity- rather than accuracy-related motivations for “fake news” sharing will help policy-makers, practitioners and research to develop new interventions that may effectively stop the proliferation of fake news on social media platforms. To enforce norms of sharing high-quality information, potential interventions should target the social and reputational costs associated with fake news sharing. One such strategy could be to remind social media users “not to make fools of themselves” when sharing information from untrustworthy sources (Altay, Hacquin and Mercier 2019). Similarly, while fact checking will likely not discourage offenders from sharing fake news, publicly tagging that a story contains falsehoods increases the reputational damage of sharing falsehoods. Although such interventions are incomplete and imperfect solutions (Smith, Jackson and Raj 2017) – fake news sharing likely reflects deep structural forces, like increased political polarization (Spohr 2017) – they have the potential to substantially increase the overall quality of information shared on social media.

Another interesting possibility is that individual-level factors shape how strong partisan and accuracy motivations are. We report a replication of Pennycook and Rand’s (2019b) analysis of partisan bias in truth discernment with each of our three samples in Section E.1 in the OA. Strikingly, although we replicate the original result and find no partisan bias in the MTurk and Twitter samples, substantial partisan bias emerges in the most diverse, Lucid sample. It is beyond the scope of the current analysis to reconcile these findings, but it is a stark reminder that psychological motivations are not set in stone and that researchers must be mindful of the samples they employ when studying “fake news”.
Materials and methods

Sample and participants

Our participants are Twitter users recruited by YouGov survey agency to be part of a multi-wave panel survey. We chose to run our study on Twitter because it is the largest and most relevant social network (in the Western world), which provides researchers sufficient access to its data to implement a field experiment on “fake news” sharing. Admittedly, the overwhelming focus on Twitter is a weakness of the “fake news” literature, but our ambition here was not to broaden the scope of our knowledge by looking at other platforms but to deepen our understanding of psychological procedures that have been studied previously. As such, we consider it a strength of our study that its findings speak directly to previous works in the literature.

We fielded the first wave of the survey between December 2018 and January 2019, where participants gave informed consent to scraping their Twitter feeds based on the Twitter usernames they disclosed and to linking it to their survey responses. We embedded the present experiment in Wave 2 of the panel, which was fielded between November 2019 and December 19, 2019. 1,870 people participated in our experiment, but for 270 respondents we could not match their self-reported usernames to a Twitter account, which prevented us from scraping their feeds and analysing their behavior. Our final sample size is therefore N = 1,600. Our Twitter sample resembles the Twitter population in terms of partisanship, income and gender, but our respondents are older, whiter and more educated than the average user (see OA2 on detailed demographic information).

Our survey first asked respondents a series of political questions, then randomly assigned them to either a treatment condition where they watched a video on how to spot “fake news” online or to a control condition where they did not watch any videos. Following a few filler questions, all respondents participated in a fake news quiz where they rated the perceived accuracy of eight news headlines.

Educational videos

We sought to find the best short educational videos publicly available. Beyond searching the internet ourselves, we solicited candidate videos through the Social Observatory for Dis-
information and Social Media Analysis (SOMA). SOMA is an international, interdisciplinary project founded by the European Commission to fight fake news with a diverse expert panel. Based on the recommendations of the experts and our own results, we compiled a list of 18 videos that fulfilled our criteria: 1) professionally produced, 2) in English, 3) under 10 minutes long, and 4) aimed at an adult audience. We chose the six best videos based on our assessment of the professionalism of its production and the specificity of recommendations for spotting fake news.

Table 1 describes the details of the six videos, which are produced by respectable governmental and non-governmental organizations, educators and media companies. The six videos combined have been viewed almost one million times only on YouTube.

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Link</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>How to choose your news -</td>
<td>Damon Brown</td>
<td>youtu.be/q-Y-z6HmRgI</td>
<td>517,000</td>
</tr>
<tr>
<td>Five ways to spot fake news</td>
<td>Quartz</td>
<td>youtu.be/y7eCB2F89K8</td>
<td>46,000</td>
</tr>
<tr>
<td>How to Spot Fake News -</td>
<td>FactCheck.org</td>
<td>youtu.be/AkwWcHekMdo</td>
<td>220,000</td>
</tr>
<tr>
<td>The Fact Checker’s guide for detecting fake news</td>
<td>Factcheck.org</td>
<td>youtu.be/SoWCDJAMk2Y</td>
<td>NA</td>
</tr>
<tr>
<td>Helping Students Identify Fake News with the Five C’s of Critical Consuming</td>
<td>Washington Post</td>
<td>youtu.be/xf8mjbVRqao</td>
<td>89,000</td>
</tr>
<tr>
<td>Four ways to tell if something is true online - Break the Fake</td>
<td>John Spencer</td>
<td>youtu.be/E-049KTrYBg</td>
<td>290</td>
</tr>
</tbody>
</table>

All six videos recommend simple techniques for establishing the veracity of online news stories such as reading beyond the headline, considering the source of the information, cross-checking with other sources. Section A.1 in the Online Appendix details specific recommendations for each educational video.

Fake news quiz

We built on previous research by Pennycook and colleagues to develop our fake news quiz, which consisted of 8 items (4 fake, 4 real; 4 pro-Republican, 4 pro-Democratic fully

---

5 The Washington Post’s video has been distributed through the newspaper’s website. We are grateful to Glenn Kessler at the Washington Post for uploading their video to YouTube to aid our research.
Each item consists of a real news headline, a lead (one-sentence summary of the news) and a picture, mimicking the display of news links on social media. The items are displayed in Figure OA1. The items were presented in a random order, and participants were asked the following question: “To the best of your knowledge, does this headline accurately describe an event that actually happened?” The answer options included 1) Definitely not, 2) Probably not, 3) Probably yes, and 4) Definitely yes.

We calculate truth discernment scores by subtracting the average accuracy ratings of fake news from the average accuracy ratings of real news and dividing it by three. The truth discernment scale has a theoretical range from -1 to 1, where 0 represents no ability to discern between fake and real news (answering as if at random), 1 represents an excellent ability to discern between fake and real news, and -1 corresponds to a person who believes every fake news story and disbelieves every real news story. Because no such person participated in our study, we focus our analyses on the range between 0 and 1.

Our objective was to rely on established measures of truth discernment and to reduce measurement error as much as possible, despite the hard space constraints in our survey. We reanalyzed open data from Pennycook et al. (2020) and ran a pilot study on MTurk (N = 800) performing item response theory models to identify the eight items with the highest discrimination score (for details see Section C in the OA).

Scraping Twitter, counting fake news

Scraping of Twitter’s tweets has been in operation since 04-03-2019 using Twitter’s API’s ”user timeline” endpoint. The scraping is conducted weekly. The tweets were collected using the free account, which limits scraping to the 3200 most recent tweets from this endpoint. Tweets much older than this timepoint are thus possibly not included.

The URLs are extracted from each tweet’s json file in the ”entities urls” section of the json file. We primarily used the software program Python’s urlexpander package (Yin 2018) to try to expand URLs to get the full domain.

To test the robustness of the the expansion of URLs, we also tried Pythons urllib and http modules, which are part of The Python Standard Library (Van Rossum and Drake 2009). This made essentially no difference in terms of fake news detected.
After the URLs were expanded, we used the urlexpander package to extract the domain from the URLs. If a URL was already expanded, we simply extracted the domain from the not expanded URL.

**Analysis strategy**

Truth discernment is modeled in a linear multilevel model with varying slopes for the six videos. While it is rare to model varying slopes without random intercepts, in our analysis the fixed intercepts simply denote that notwithstanding the video, all participants in the control condition have the same level of discernment.

We aggregate our Twitter data to user-weeks such that for each respondent, the day of the experiment is coded as 0, and seven-day periods going into the future and the past are counted. For example, if our user-week variable is at 1, it captures the user’s behavior the seven days after the experiment; if it is at -3, it captures behavior in the third week before exposure.

“Fake news” sharing is modelled in difference-in-differences models that regress the outcome on a dummy for being in the treatment group, a dummy that switches on the day after the respondent participated in our study, and an interaction term between the two variables.

We model number of “fake news” stories shared in negative binomial models. We model the binary outcome (did the respondent share any fake news in a given week?) and the proportion of fake to real news in binomial regression models. Finally, we model the news source quality in a simple linear regression. We run all models on “fake news” sharing in a Bayesian setting relying on “stan” and the “brms” R package. We defined weakly informative priors centered at 0, which do not affect our estimates but help with model convergence.

**Note on pre-registrations**

We pre-registered our hypotheses and analyses both with regards to the causal effects of the fact-checking videos on truth discernment and fake news sharing, and for the relationship between truth discernment and fake news sharing. Our pre-registrations are posted at the end of the online appendix.

We depart from our pre-registrations in the following ways: 1) we excluded a single respondent in the control group who shared a very large number of fake news stories in the
pre-treatment period but then substantially reduced their fake news disseminating activity a few weeks before the experiment was fielded. This exclusion made our count model estimates more conservative. 2) Our models regressing fake news sharing on truth discernment include an extra covariate for being in the treatment group because – at least in theory – being exposed to educational videos could affect both truth discernment and fake news sharing. Dropping this covariate from our models or relying on fake news sharing in the pre-treatment period does not change any of our findings. 3) Whereas we planned to run multilevel models investigating the treatment effect for each of the six educational videos, we refrained from this in the analysis involving “fake news” sharing because the models had trouble converging and provided very random effect noisy estimates.
References


Altay, Sacha, Anne-Sophie Hacquin and Hugo Mercier. 2019. “Sharing Fake News is Bad for Your Epistemic Reputation.”


EC. 2018. “Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Action Plan against Disinformation (JOIN(2018) 36 final).”


URL: https://doi.org/10.1073/pnas.1920498117


Osmundsen, Mathias, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann and Michael Bang Petersen. 2020. “Partisan polarization is the primary psychological motivation behind “fake news” sharing on Twitter.”. URL: https://doi.org/10.31234/osf.io/v45bk


URL: https://www.pnas.org/content/116/7/2521


URL: psyarxiv.com/6m4ts


Siegel, A and Vivienne Badaan. 2020. “#No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online.” Forthcoming in the American Political Science Review.


URL: http://arxiv.org/abs/1704.07506


URL: https://doi.org/10.5281/zenodo.1345144

23
Online Appendix

Contents

A Experimental materials
  A.1 Specific recommendations by educational videos . . . . . . . . . . . . . . . . . . . 25
  A.2 Fake news quiz . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26

B Descriptive statistics
  B.1 Twitter sample demographics . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28
  B.2 Fake news sharing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

C Lucid pilot fine-tuning the fake news quiz 30

D Model details
  D.1 Educational videos and truth discernment . . . . . . . . . . . . . . . . . . . . . . . 33
  D.2 Disaggregating discernment into gullibility and trust in mainstream media . . . 34
  D.3 Truth discernment and news sharing . . . . . . . . . . . . . . . . . . . . . . . . . 35

E Additional (post-hoc) analyses
  E.1 Partisan bias in truth discernment and sharing . . . . . . . . . . . . . . . . . . . . 38
A Experimental materials

A.1 Specific recommendations by educational videos

“How to choose your news” - TED-Ed

1. Get original news cutting out the middlemen
2. In chaotic times, do not follow news constantly
3. Check multiple sources
4. Be aware of anonymous sources
5. Verify before spreading

“Five ways to spot fake news” - Quartz

1. Consider where the information is coming from
2. Consider if the headline sound neutral
3. Consider who wrote it
4. Consider what the resources are
5. Consider if the images are accurate

“How to Spot Fake News” - Factcheck.org

1. Consider the source
2. Read beyond headline
3. Check the author
4. Consider what the support is for a claim
5. Check the date
6. Consider if it could be satire
7. Check biases
8. Consult experts

“The Fact Checker’s guide for detecting fake news” - Washington Post

1. Double check the url
2. Consider if the photo seems unrealistic
3. Check the sources
4. Use dedicated plug-ins

“Fake News with the Five C’s of Critical Consuming” - John Spencer

1. Context
2. Credibility
3. Construction
4. Corroboration
5. Compare

“Four ways to tell if something is true online - Break the Fake” - MediaSmarts

1. Use fact checking tools
2. Find the source
3. Verify the source
4. Check other sources

A.2 Fake news quiz

Participants saw the following instructions prior to starting the fake news quiz:

In the next section, we will present you 8 headlines from news stories that appeared on the internet. Your task will be to indicate to the best of your knowledge, whether each headline accurately describes an event that actually happened on the following scale: 1 (definitely not), 2 (probably not), 3 (probably yes), and 4 (definitely yes).

Note that we are NOT interested whether you agree with a claim someone (allegedly) made. Instead, we would like to know whether you think the claim was actually made by that person. Similarly, we are NOT interested whether you are happy about an event that (allegedly) happened. Instead, we want to know whether you think that the event actually happened or not.

Please note that the images may take a moment to load.
<table>
<thead>
<tr>
<th></th>
<th>Fake news</th>
<th>Real news</th>
</tr>
</thead>
</table>
| **Pro-Dem.**              | ![Image](image1.png) **Journalist Who Filmed Trump Sobbing in Rose Garden Still in Custody**  
WASHINGTON, D.C. - A freelance photjournalist remains in custody nearly a week after he was detained by officers from the Secret Service's Uniformed...  
REALNEWSRIGHTNOW.COM | ![Image](image2.png) **New plagiarism row for Melania Trump as pamphlet bears uncanny likeness to 2014 version**  
Booklet on children's piano safety for the lady in the lead campaign seems almost identical to one first published by FTC two years ago.  
THEGUARDIAN.COM |
| **Pro-Dem.**              | ![Image](image3.png) **Trump: “Why would I care about the climate? I’ll be dead in 10 years anyway”**  
Donald Trump has once again managed to insult the international community. The US is pulling out of the Paris climate agreement ends 20 years of tough negotiations. Now, in an...  
THEPOSTILLON.COM | ![Image](image4.png) **Trump once said pleading the Fifth was for ‘the mob’ and now he might do it**  
Donald Trump has denied the Fifth Amendment as the refuge of mobsters, and Richard Nixon infamously told aides, “I want you all to stonewall it, let them plead the Fifth Amendment...  
MON.COM |
| **Pro-Rep.**              | ![Image](image5.png) **Obama Stole $1 BILLION From Social Security, Gave It To THIS Group!**  
Unlawful immigration. Controlling immigration is not an issue of such a large number of different issues in our nation. Among these issues are violations like medical or trafficking, human...  
USACONSERVATIVEREPORT.COM | ![Image](image6.png) **Darryl Strawberry says Donald Trump ‘is a great man’**  
Darryl Strawwoery may want athletes to ‘stick to sports,’ but clearly he sees himself as the exception.  
NIKEXNEWS.COM |
| **Pro-Rep.**              | ![Image](image7.png) **President Trump Fires All 14 Muslim Federal Judges**  
And for good reason.  
AMERICANSAMISRESEERG.ORG | ![Image](image8.png) **Massachusetts Democrat charged in bribery, kickback schemes**  
A former state senator from Massachusetts was charged last week with collecting about $1 million in bribes and kickbacks that he allegedly laundered through his company.  
FOREX.COM |
B Descriptive statistics

B.1 Twitter sample demographics

Table OA2 displays the demographic characteristics of our sample. It demonstrates first that the two treatment groups are similar in terms of demographics. It also displays the PEW Research Center’s estimates of the U.S. Twitter population. It shows that our sample is somewhat older, more white, and more educated than the Twitter population, while it is similar in terms of gender, income and partisan identity.

Table OA2: Sample characteristics by treatment group

<table>
<thead>
<tr>
<th></th>
<th>Control group</th>
<th>Video group</th>
<th>PEW benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age: 18-29</td>
<td>0.10</td>
<td>0.13</td>
<td>0.29</td>
</tr>
<tr>
<td>Age: 30-49</td>
<td>0.46</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>Age: 50-64</td>
<td>0.33</td>
<td>0.31</td>
<td>0.19</td>
</tr>
<tr>
<td>Age: 65+</td>
<td>0.11</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>Black</td>
<td>0.08</td>
<td>0.08</td>
<td>0.11</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.09</td>
<td>0.10</td>
<td>0.17</td>
</tr>
<tr>
<td>White</td>
<td>0.78</td>
<td>0.75</td>
<td>0.60</td>
</tr>
<tr>
<td>No high school</td>
<td>0.01</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>High school</td>
<td>0.31</td>
<td>0.31</td>
<td>0.54</td>
</tr>
<tr>
<td>College+</td>
<td>0.68</td>
<td>0.68</td>
<td>0.42</td>
</tr>
<tr>
<td>Female</td>
<td>0.52</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>&lt;30K USD</td>
<td>0.18</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td>30K-75K USD</td>
<td>0.41</td>
<td>0.39</td>
<td>0.36</td>
</tr>
<tr>
<td>75K+ USD</td>
<td>0.41</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>Republican</td>
<td>0.29</td>
<td>0.27</td>
<td>0.35</td>
</tr>
<tr>
<td>Democrat</td>
<td>0.59</td>
<td>0.61</td>
<td>0.60</td>
</tr>
<tr>
<td>Independent</td>
<td>0.12</td>
<td>0.12</td>
<td>NA</td>
</tr>
<tr>
<td>N</td>
<td>799</td>
<td>801</td>
<td></td>
</tr>
</tbody>
</table>

---

"Sizing up Twitter users" by Stefan Wojcik and Adam Hughes https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ last accessed 2020-06-06
B.2 Fake news sharing

Figure OA1 displays the four outcome variables tapping into fake news sharing on Twitter. It shows that particularly the raw “fake news” counts are rather noisy, especially in the treatment group (see panel A). Meanwhile, dichotomizing the measure (does respondent shares at least one “fake news” on this week? - panel B) and looking at mean news quality (panel D), reduce the noise. Remember that even though we are running difference-in-differences models, our identification strategy is rooted in the randomised treatment, therefore usual concerns about parallel trends do not concern us.

Figure OA1: Fake news sharing in the treatment and control group by time

Note: Points indicate the sum or mean value in the experimental group. Lines are loess curves.
C Lucid pilot fine-tuning the fake news quiz

Developing a precise measure of truth discernment ability has been an important challenge for this study, as it is featured prominently both in the analyses of the video intervention and the relationship between sharing and believing “fake news”. While ideally we could have relied on a long battery with dozens of items to get a nuanced measure, this was not possible in the Twitter sample due to limits in respondent time (and more broadly, resources). Instead, we had to make sure that the items we rely on offer the least noise estimates possible. Thus, we ran a pilot study on Lucid, a large marketplace for online survey panels, relying on a sample of 997 US respondents quota sampled to resemble the marginal distribution of population demographics in terms of age, gender, education, race and region.

The ambition of this pilot was to get as much data on the performance of “fake news” quiz items as possible. Therefore, we took the largest publicly available inventory of quiz items containing 14 Fake Pro-Democratic, 26 Fake Pro-Republican, 15 Real Pro-Democratic, and 20 Real Pro-Republican items (Pennycook et al. 2020). We invited participants in our pilot to rate the accuracy of 32 items each, randomly sampling 8 items from each of the 4 categories, presented in a random order.

We employ a simple, two parameter item response theory (IRT) model to identify the best items tapping into the latent trait of truth discernment. IRT is a psychometric method developed for designing and improving tests. The two parameter model estimates an item difficulty parameter i.e. how easy or hard an item is, and a discrimination parameter i.e. how good the item is at distinguishing between people with similar ability. As a rule of thumb, the best tests have items with varying difficulty and high discrimination.

Importantly, IRT models assume that the underlying latent construct is uni-dimensional. Truth discernment does not satisfy this criterion, because dispositions such as cynicism (a general disbelief in all news) will have opposite effects on truth discernment when it comes to “fake news” items (cynicism improves performance) and “real news” items (cynicism undermines performance). Therefore, we analyse each of our 4 categories separately. Although participants answer the items on a 4 point scale (as described in Section A.2), we transformed these into a dummy of correctly / incorrectly perceiving news accuracy.

Ideally, all participants would have rated all items, but Lucid poses a 15 minute time limit on surveys and respondents may have experienced fatigue sooner than that.
Figure OA2: Truth discernment item difficulty and discrimination parameter estimates from a 2-factor IRT model

Note: Error bars denote standard errors. Blue dots mark the parameter estimates for the 8 (+ 1) items, which are subsequently employed in our MTurk and Twitter samples.

Figure OA2 summarises our findings plotting item difficulty (x axis) against discrimination (y axis) along with error bars denoting standard errors. Several conclusions emerge. First, all fake news items and pro-Republican real news items show high variance both on difficulty and discrimination. Although the data is too noisy to be sure that the best items are significantly better than the most other items at 95%, they are efficient items. Second, real pro-Democratic items are much more difficult and on average substantially worse at discriminating between underlying ability. Shedding light on this interesting asymmetry is an interesting avenue for future research.

The attentive reader may notice that 3 items are highlighted with blue in the Real Pro-Republican panel. This is because one of the items (“Foxnews.com: Deep state has ‘weaponized’
security clearances against Trump, conservative Pentagon official’s lawyer says”) has proved to be an outlier in the MTurk pilot, and has been swapped for the next best item (“NYDailynews.com: Darryl Strawberry says Donald Trump ‘is a great man’).
D Model details

D.1 Educational videos and truth discernment

**Table OA3:** Main effects from MLM models

<table>
<thead>
<tr>
<th>Dependent variable:</th>
<th>MTurk pilot</th>
<th>Twitter sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth discernment</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>ATE</td>
<td>0.07*</td>
<td>0.05**</td>
</tr>
<tr>
<td></td>
<td>(0.02, 0.13)</td>
<td>(0.02, 0.08)</td>
</tr>
<tr>
<td>Control group</td>
<td>0.30***</td>
<td>0.37***</td>
</tr>
<tr>
<td></td>
<td>(0.25, 0.35)</td>
<td>(0.35, 0.39)</td>
</tr>
<tr>
<td>Observations</td>
<td>686</td>
<td>1,600</td>
</tr>
</tbody>
</table>

*Note:* *p<0.05; **p<0.01; ***p<0.001

**Table OA4:** Random effects for the six videos from MLM models

<table>
<thead>
<tr>
<th>Group</th>
<th>ATE</th>
<th>SE</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Pooled estimate</td>
<td>0.07</td>
<td>0.03</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>2 FactCheck.org</td>
<td>0.06</td>
<td>0.02</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>3 John Spencer</td>
<td>0.08</td>
<td>0.02</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>4 MediaSmarts</td>
<td>0.10</td>
<td>0.02</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>5 Quartz</td>
<td>0.09</td>
<td>0.02</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>6 TED-Ed</td>
<td>0.04</td>
<td>0.02</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>7 Washington Post</td>
<td>0.07</td>
<td>0.02</td>
<td>MTurk pilot</td>
</tr>
<tr>
<td>8 Pooled estimate</td>
<td>0.05</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
<tr>
<td>9 FactCheck.org</td>
<td>0.04</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
<tr>
<td>10 John Spencer</td>
<td>0.06</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
<tr>
<td>11 MediaSmarts</td>
<td>0.06</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
<tr>
<td>12 Quartz</td>
<td>0.02</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
<tr>
<td>13 TED-Ed</td>
<td>0.07</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
<tr>
<td>14 Washington Post</td>
<td>0.07</td>
<td>0.02</td>
<td>Twitter sample</td>
</tr>
</tbody>
</table>
## D.2 Disaggregating discernment into gullibility and trust in mainstream media

**Table OA5:** Main effects from MLM models

<table>
<thead>
<tr>
<th></th>
<th>Gullibility</th>
<th>Trust in MSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mturk pilot (1)</td>
<td>Mturk pilot (3)</td>
</tr>
<tr>
<td>ATE</td>
<td>−0.04 (−0.09, 0.01)</td>
<td>0.03 (−0.01, 0.07)</td>
</tr>
<tr>
<td></td>
<td>−0.04*** (−0.07, −0.02)</td>
<td>0.01 (−0.01, 0.03)</td>
</tr>
<tr>
<td>Control group</td>
<td>0.26*** (0.21, 0.30)</td>
<td>0.56*** (0.53, 0.59)</td>
</tr>
<tr>
<td></td>
<td>0.24*** (0.23, 0.26)</td>
<td>0.61*** (0.60, 0.62)</td>
</tr>
<tr>
<td>Observations</td>
<td>686</td>
<td>686</td>
</tr>
<tr>
<td></td>
<td>1,600</td>
<td>1,600</td>
</tr>
</tbody>
</table>

*Note:*  *p<0.05; **p<0.01; ***p<0.001*
D.3 Truth discernment and news sharing

Here we report full models for the relationship between truth discernment and fake fake news sharing.

**Table OA6:** Negative binomial regression of # of fake news shared on discernment and co-variates

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.17</td>
<td>-0.87*</td>
<td>-3.61*</td>
</tr>
<tr>
<td></td>
<td>[-0.51; 0.94]</td>
<td>[-1.69; -0.01]</td>
<td>[-5.59; -1.69]</td>
</tr>
<tr>
<td>Truth Discernment</td>
<td>-0.85</td>
<td>-0.38</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>[-2.30; 0.55]</td>
<td>[-1.73; 0.97]</td>
<td>[-1.55; 1.41]</td>
</tr>
<tr>
<td>Video</td>
<td>0.89*</td>
<td>0.49</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>[0.35; 1.44]</td>
<td>[-0.10; 1.06]</td>
<td>[-0.20; 0.91]</td>
</tr>
<tr>
<td>Age</td>
<td>3.81*</td>
<td>4.10*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2.49; 5.18]</td>
<td>[2.70; 5.49]</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>-0.09</td>
<td>-0.72*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.68; 0.48]</td>
<td>[-1.34; -0.15]</td>
<td></td>
</tr>
<tr>
<td>Higher education</td>
<td>-0.79*</td>
<td>-0.69*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-1.56; -0.08]</td>
<td>[-1.40; -0.05]</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>-0.29</td>
<td>-0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-1.04; 0.39]</td>
<td>[-1.52; 0.02]</td>
<td></td>
</tr>
<tr>
<td>Party ID</td>
<td>3.03*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2.24; 3.87]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRT</td>
<td></td>
<td>-0.10</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[-0.37; 0.17]</td>
<td></td>
</tr>
<tr>
<td>Trolling</td>
<td></td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[-0.33; 0.38]</td>
<td></td>
</tr>
<tr>
<td>Cynicism</td>
<td></td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[-0.10; 0.43]</td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td></td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[-0.04; 0.52]</td>
<td></td>
</tr>
</tbody>
</table>

* Null hypothesis value outside 95% credible interval (reported in square brackets).
Table OA7: Binomial regression of 1+ fake news shared on discernment and covariates

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-2.08^*$</td>
<td>$-3.12^*$</td>
<td>$-3.74^*$</td>
</tr>
<tr>
<td></td>
<td>[-2.40; -1.76]</td>
<td>[-3.71; -2.56]</td>
<td>[-4.97; -2.54]</td>
</tr>
<tr>
<td>Truth Discernment</td>
<td>0.15</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>[-0.50; 0.78]</td>
<td>[-0.49; 0.90]</td>
<td>[-0.64; 0.99]</td>
</tr>
<tr>
<td>Video</td>
<td>0.00</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>[-0.30; 0.30]</td>
<td>[-0.28; 0.33]</td>
<td>[-0.25; 0.38]</td>
</tr>
<tr>
<td>Age</td>
<td>2.39*</td>
<td>2.40*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[1.64; 3.13]</td>
<td>[1.59; 3.25]</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.03</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.28; 0.34]</td>
<td>[-0.42; 0.25]</td>
<td></td>
</tr>
<tr>
<td>Higher education</td>
<td>-0.13</td>
<td>-0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.46; 0.20]</td>
<td>[-0.47; 0.21]</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>-0.02</td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.41; 0.38]</td>
<td>[-0.47; 0.36]</td>
<td></td>
</tr>
<tr>
<td>Party ID</td>
<td>0.49*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.05; 0.93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRT</td>
<td>0.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.11; 0.18]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trolling</td>
<td>-0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.24; 0.15]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cynicism</td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.09; 0.23]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.10; 0.20]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Null hypothesis value outside 95% credible interval (reported in square brackets).
Table OA8: Negative binomial regression of # of real news shared on discernment and covariates

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.28*</td>
<td>0.75*</td>
<td>-0.39</td>
</tr>
<tr>
<td></td>
<td>[1.93; 2.67]</td>
<td>[0.17; 1.37]</td>
<td>[-1.48; 0.73]</td>
</tr>
<tr>
<td>Truth Discernment</td>
<td>2.80*</td>
<td>2.82*</td>
<td>2.01*</td>
</tr>
<tr>
<td></td>
<td>[2.12; 3.48]</td>
<td>[2.22; 3.40]</td>
<td>[1.32; 2.69]</td>
</tr>
<tr>
<td>Video</td>
<td>-0.02</td>
<td>-0.25</td>
<td>-0.00</td>
</tr>
<tr>
<td></td>
<td>[-0.33; 0.30]</td>
<td>[-0.57; 0.06]</td>
<td>[-0.32; 0.33]</td>
</tr>
<tr>
<td>Age</td>
<td>3.68*</td>
<td>3.51*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2.94; 4.42]</td>
<td>[2.77; 4.26]</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.16</td>
<td>-0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.16; 0.48]</td>
<td>[-0.55; 0.12]</td>
<td></td>
</tr>
<tr>
<td>Higher education</td>
<td>0.08</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.26; 0.41]</td>
<td>[-0.61; 0.06]</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>-0.40</td>
<td>-0.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.83; 0.03]</td>
<td>[-0.62; 0.15]</td>
<td></td>
</tr>
<tr>
<td>Party ID</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.39; 0.59]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRT</td>
<td>-0.38*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.50; -0.24]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trolling</td>
<td>0.28*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.11; 0.45]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cynicism</td>
<td>-0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.19; 0.11]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge</td>
<td>0.52*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.39; 0.65]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Null hypothesis value outside 95% credible interval (reported in square brackets).
E Additional (post-hoc) analyses

E.1 Partisan bias in truth discernment and sharing

As further evidence for the important theoretical distinction between truth discernment and fake news sharing, here we replicate Pennycook and Rand’s 2019 finding that partisan bias does not taint performance in the fake news quiz. In an influential paper called “Lazy not biased”, Pennycook and Rand argue that low cognitive reflection but not partisan motivated reasoning is responsible for believing fake news. They argue that if partisans believed all news congruent with their identity, this would undermine their ability to perform well at the fake news quiz. The stronger a respondent identifies with Republicans, the worse they will perform at discerning between pro-Republican fake and real news, and vice versa for Democrats. To test whether there is such a partisan asymmetry in truth discernment, we calculate for each respondent a separate discernment score for pro-Republican and pro-Democratic news. Next, we regress these scores on a binary indicator for source partisanship, a classic seven-point continuous partisan identity scale and the interaction of these two terms.

Figure OA3: Truth discernment by partisan alignment

Figure OA3 plots identification with the Republican party (x axis), against truth discern-
ment scores (y axis) split by news source leaning (colors). Recall, if partisan motivation made people gullible to congruent news, we would see a descending red line (stronger Republican identification leading to worse pro-Republican discernment), and an ascending blue line. In contrast, Figure OA3 shows a horizontal red line (discernment between pro-Republican fake and real news is similar across the range partisan identification), and a descending blue line: Strong Democrats are actually the best at discerning between pro-Democratic fake and real news. Strong Republicans appear to perform worse here, because they disbelieve pro-Democratic real news.

**Figure OA4:** Mean perceived accuracy of fake and real news headlines (as a function of political concordance)

---

Pennycook and Rand (2019) present the same substantive finding slightly differently, by calculating the mean perceived accuracy for politically concordant and discordant fake and real news. Thus, Republican respondents’ ratings of pro-Republican news and Democrats’ ratings of pro-Democratic news contribute to concordant ratings, whereas Republicans’ ratings of pro-Democratic news, and Democrats’ ratings of pro-Republican news contribute to discordant ratings. We plot these group averages in Figure OA4, directly replicating Figure 3 on page 43 in the original publication. Although respondents on average believe all politically discordant news slightly less than concordant news, there is no evidence for a partisan bias.

In short, our analyses reproduces the finding that there is no partisan bias in believes in
fake news in contemporary American society. Importantly, wherein the original study relied on convenience samples via Amazon’s MTurk, our results lead to the same conclusions based on a sample of verified Twitter users. These findings stand in stark contrast with recent findings showing that strong Republicans are much more likely to share pro-Republican fake news than Democrats (e.g. Osmundsen et al. [2020]).
Truth discernment and sharing of fake and real news on Twitter (#31178)

1) Have any data been collected for this study already?
It’s complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What’s the main question being asked or hypothesis being tested in this study?
H1a. The higher a person’s ability to discern fake news from real news, the fewer fake news stories they will share on Twitter.
H1b. The higher a person’s ability to discern fake news from real news, the more fake news stories they will share on Twitter.
H2. The higher a person’s ability to discern fake news from real news, the more real news stories they will share on Twitter.

3) Describe the key dependent variable(s) specifying how they will be measured.
The key DV is the number of fake news stories shared on Twitter from verified fake news websites in the period between 2018 November (12 months prior to data collection) and 2020 February (3 months after data collection). The verified fake news website list comes from Guess, Andrew, Jonathan Nagler, and Joshua Tucker. “Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook.” Science Advances 5, no. 1 (2019). https://doi.org/10.1126/sciadv.aau4586.

4) How many and which conditions will participants be assigned to?
This is a cross-sectional study. Participants will be participating in two experiments unrelated to the present pre-registration.
Our main independent variable is truth discernment. Participants take a “fake news quiz” seeing 8 news headlines, (4 fake and 4 real). They are asked to rate each headline whether it accurately describes an event that actually happened, truth discernment is calculated as the average accuracy ratings of real news minus average accuracy ratings of fake news divided by 3. Higher values indicate better discernment.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Our main analysis concerns the correlation between the number of fake news shared and the truth discernment score. We will run three regression models, regressing number of fake news shared on truth discernment score 1) without any covariates, 2) with demographic covariates (age, gender, and education), 3) with psychological covariates (partisan identity, cognitive reflection score, trolling, political cynicism, and political knowledge).

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We exclude participants who provided two different yet valid Twitter ID in Wave 1 and Wave 2.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
This is wave 2 of a panel study of Twitter users, who consented to scrape their feeds. In the first wave, 3016 participants gave their consent by entering a Twitter ID in the appropriate field. Of these, we could identify and scrape 2336 users. All 3016 participants will be invited to participate in the survey, but naturally we can include in the study only participants who provided a valid Twitter ID in at least one of the two surveys.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Truth discernment will be measured with a quiz implemented in wave 2 of our panel study, which is yet to be run. However, we will rely on fake news counts from participants’ scraped Twitter feed. Parts of this data we already analyzed for a different study.

We consider it useful to distinguish between three periods when it comes to participants’ Twitter behavior. 1) prior to enrolling in wave 1 of our study (~2019 February), 2) period between the two waves (2019 Feb – 2019 Nov), 3) period following Wave 2 (2019 Nov – ). Experimental demand effects and Twitter’s efforts to reduce fake news on the network affect each period differently. Consequently, we will run robustness checks splitting the DV to these three periods.
The individual-level causal effects of popular anti-“fake news” videos (#31174)

1) Have any data been collected for this study already?
It’s complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What’s the main question being asked or hypothesis being tested in this study?
A) Does exposure to an educational anti fake news video improve truth discernment (the ability to distinguish between fake news and real news)?
B) Does exposure to an educational anti fake news video reduce the number of fake news shared on Twitter?

3) Describe the key dependent variable(s) specifying how they will be measured.
A) Participants will take a “fake news quiz”, seeing 8 news headlines, (4 fake and 4 real). They are asked to rate each headline whether it accurately describes an event that actually happened using the following scale: 1 (definitely not), 2 (probably not), 3 (probably yes), and 4 (definitely yes).
Our DV is the truth discernment score i.e. the average accuracy ratings of real news minus average accuracy ratings of fake news divided by 3. Higher values indicate better discernment.
B) The number of links shared from verified fake news websites in the 3 months before and after treatment.

4) How many and which conditions will participants be assigned to?
Control: No video. 50% of sample
Treatment: Asked to watch an educational videos on how to spot fake news. 50% of sample. Participants in the treatment group are randomly allocated to one of six videos. All videos are available on Youtube.
1. Four ways to tell if something is true online – MediaSmarts, https://youtu.be/E-049KTrYBg,

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
A) Simple independent samples t-test (control vs treatment) with truth discernment scores.
B) Difference-in-differences estimator of the number of fake news stories shared on Twitter with 3 months prior to exposure as the baseline period and 3 months after the exposure as the treatment period.

For both analyses we will be running multilevel models estimating the effect of the specific videos.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
We exclude participants who fail an audio/video equipment test or provide two different yet valid Twitter ID in Wave 1 and Wave 2 of our survey.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined.
This is wave 2 of a panel study of Twitter users, who consented to share data from their feeds. In the first wave, 3016 participants gave their consent by entering a Twitter ID in the appropriate field. Of these, we could identify and scrape 2336 users. All 3016 participants will be invited to participate in the survey, but naturally we can include in the study only participants who provided a valid Twitter ID in at least one of the two surveys.

8) Anything else you would like to pre-register? (e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We already possess, although have not analyzed the data constituting the baseline of fake news sharing behavior. We have *not* run the experiment yet. We will run robustness checks varying the size of the window around the treatment to identify fake news sharing prior and after participant in the experiment.

Available at https://aspredicted.org/blind.php?x=gv3cf8