A digital media literacy intervention increases discernment between mainstream and false news in the United States and India

Andrew M. Guess[a,1,2], Michael Lerner[b,c,1], Benjamin Lyons[d,1], Jacob M. Montgomery[e,1], Brendan Nyhan[f,1], Jason Reifler[g,1], and Neelanjan Sircar[h,1]

[a]Department of Politics, Princeton University, Princeton, NJ 08544; [b]Department of Political Science, University of Michigan, Ann Arbor, MI 48109-1045; [c]Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, MI 48109-1045; [d]Department of Communication, University of Utah, Salt Lake City, UT 84112; [e]Department of Political Science, Washington University in St. Louis, St. Louis, MO 63130-4899; [f]Department of Government, Dartmouth College, Hanover, NH 03755; [g]Department of Politics, University of Exeter, Exeter EX4 4RJ, United Kingdom; and [h]Department of Political Science, Ashoka University, Sonipat, Haryana 131029, India

Edited by David G. Rand, Massachusetts Institute of Technology, Cambridge, MA, and accepted by Editorial Board Member Margaret Levi April 28, 2020 (received for review November 20, 2019)

Widespread belief in misinformation circulating online is a critical challenge for modern societies. While research to date has focused on psychological and political antecedents to this phenomenon, few studies have explored the role of digital media literacy shortfalls. Using data from preregistered survey experiments conducted around recent elections in the United States and India, we assess the effectiveness of an intervention modeled closely on the world's largest media literacy campaign, which provided "tips" on how to spot false news to people in 14 countries. Our results indicate that exposure to this intervention reduced the perceived accuracy of both mainstream and false news headlines, but effects on the latter were significantly larger. As a result, the intervention improved discernment between mainstream and false news headlines among both a nationally representative sample in the United States (by 26.5%) and a highly educated online sample in India (by 17.5%). This increase in discernment remained measurable several weeks later in the United States (but not in India). However, we find no effects among a representative sample of respondents in a largely rural area of northern India, where rates of social media use are far lower.

digital literacy | social media | misinformation

Social media platforms have proved to be fertile ground for inflammatory political misinformation. People around the world increasingly worry that so-called "fake news" and other forms of dubious or false information are misleading voters—a fear that has inspired government actions to address the problem in a number of countries (1, 2).

Research into online misinformation has thus far focused on political, economic, and psychological factors (3–5). In this article, we focus on another human vulnerability to online political misinformation: shortfalls in digital media literacy.

While largely overlooked in the emerging empirical literature on digital disinformation and fake news, the concept of digital media literacy usefully captures the skills and competencies needed to successfully navigate a fragmented and complex information ecosystem (6). Even under ideal conditions, most people struggle to reliably evaluate the quality of information they encounter online because they lack the skills and contextual knowledge required to effectively distinguish between high- and low-quality news content.

The connection between digital media literacy and misinformation was identified early by theorists. "Misinformation—and disinformation—breeds as easily as creativity in the fever-swamp of personal publishing," according to an influential 1997 introduction to the subject. "It will take all of the critical skills users can muster to separate truth from fiction" (ref. 7, p. xii).

More than 20 y later, these warnings seem prescient. Survey research shows that few people are prepared to effectively navigate the digital world. For example, the Pew Research Center found as recently as 2017 that only 17% of US adults have the skills and confidence to learn new information effectively online (8). Nonetheless, people worldwide increasingly obtain news and information from social media platforms that lack traditional editorial controls (9, 10), allowing politicians and other actors to widely disseminate misinformation via algorithmic news feeds. Without the necessary digital media literacy skills, people frequently fall victim to dubious claims they encounter in this context.

Significance

Few people are prepared to effectively navigate the online information environment. This global deficit in digital media literacy has been identified as a critical factor explaining widespread belief in online misinformation, leading to changes in education policy and the design of technology platforms. However, little rigorous evidence exists documenting the relationship between digital media literacy and people's ability to distinguish between low- and high-quality news online. This large-scale study evaluates the effectiveness of a real-world digital media literacy intervention in both the United States and India. Our largely encouraging results indicate that relatively short, scalable interventions could be effective in fighting misinformation around the world.

These concerns have become especially salient in the United States and India in recent years. In the United States, low-quality online articles were distributed widely on social media in the months before the 2016 US presidential election (11). This phenomenon created widespread fears that fake news was misleading people at a massive scale (12). Smartphone use has also made India, the world's largest democracy, a fertile environment for online rumors and misinformation. Viral misinformation

Author contributions: A.M.G., M.L., B.L., J.M.M., B.N., J.R., and N.S. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.G.R. is a guest editor invited by the Editorial Board.

Data deposition: Data files and scripts necessary to replicate the results in this article are available at the Dataverse repository at https://doi.org/10.7910/DVN/Q5QINN. The US study preanalysis plan is available at https://osf.io/u3sgc. The India study preanalysis plan is available at https://osf.io/97rnz.

[1] A.M.G., M.L., B.L., J.M.M., B.N., J.R., and N.S. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: aguess@princeton.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1920498117/-/DCSupplemental.

First published June 22, 2020.

spread via WhatsApp in India has reportedly provoked hatred and ethnic violence (13). Moreover, online political misinformation became a significant concern during the 2019 Indian general election as political parties engaged in aggressive digital campaign efforts via short message service (SMS) and messaging applications like WhatsApp (14, 15). For instance, one analysis found that over 25% of the news shared on Facebook during the election by the governing Bharatiya Janata Party (BJP) came from dubious outlets (16).

Many nonprofits and governments are seeking to counter these trends (and the related threat of foreign manipulation campaigns) by improving the digital media literacy of news consumers (17–20). For instance, American universities increasingly teach media literacy to undergraduate students (21) and similar efforts are also being proposed at the kindergarten to grade 12 (22). Similarly, WhatsApp and the National Association of Software and Service Companies announced plans to train nearly 100,000 people in India through in-person events and posts on social media to spot misinformation (23).

Despite the attention and resources these initiatives have received, however, little large-scale evidence exists on the effectiveness of promoting digital media literacy as a response to online misinformation. Existing scholarly work related to digital and media literacy is frequently qualitative in nature or focused on specific subpopulations and/or issues. Observational findings are mixed (24, 25) and randomized controlled trials remain rare (26).

Two related but more specific approaches have been shown to be somewhat effective in countering misinformation and are important to note, however. First, inoculation interventions have been employed to protect audiences against misleading content by warning of misinformation and either correcting specific false claims or identifying tactics used to promote it. This approach has been shown to reduce the persuasiveness of misinformation in specific domains (27–32). In addition, other studies evaluate the effectiveness of providing warnings about specific misinformation (33, 34).

We therefore seek to determine whether efforts to promote digital media literacy can improve respondents' ability to correctly evaluate the accuracy of online content across issues. Such a finding would suggest that digital media literacy shortfalls are a key factor in why people fall victim to misinformation. In particular, we consider the effects of exposure to Facebook's "Tips to Spot False News," which were developed in collaboration with the nonprofit First Draft and subsequently promoted at the top of users' news feeds in 14 countries in April 2017 and printed in full-page newspaper advertisements in the United States, the United Kingdom, France, Germany, Mexico, and India (35–40). A variant of these tips was later distributed by WhatsApp (a Facebook subsidiary) in advertisements published in Indian and Pakistani newspapers in 2018 (41, 42). These tips are therefore almost surely the most widely disseminated digital media literacy intervention conducted to date. (The full treatments are provided in *SI Appendix*, section A.) The US treatment, which was adapted verbatim from Facebook's campaign, consists of 10 strategies that readers can use to identify false or misleading stories that appear on their news feeds, whereas the India treatment, which uses adapted versions of messages shown in India by Facebook and WhatsApp, presents 6.

These interventions provide simple rules that can help individuals to evaluate the credibility of sources and identify indicators of problematic content without expending significant time or attention. For instance, one sample tip recommends that respondents "[b]e skeptical of headlines," warning that "If shocking claims in the headline sound unbelievable, they probably are." Such an approach should reduce reliance on low-effort processes that frequently lead people astray (e.g., perceptions of cognitive fluency) by teaching people more effective heuristics (e.g., skepticism toward catchy headlines). Importantly, the success of this approach does not require readers to take burdensome steps like conducting research or thinking deeply about each piece of news they encounter (which is typically impossible in practice given the volume of stories that social media users encounter). Instead, this intervention aims to provide simple decision rules that help people distinguish between mainstream and false news, which we call "discernment" following ref. 4.

There are important reasons to be skeptical about the effectiveness of this approach. Prior research has found that media literacy interventions like this can help people think critically about the media content they receive (43). However, prior studies focus mostly on offline health behavior; the extent to which these interventions are effective for controversial political claims or online (mis)information is largely unknown. Moreover, such interventions may struggle to overcome people's reliance on heuristics such as familiarity and congeniality that news consumers use to evaluate the credibility of online stories (44, 45). Finally, attempting to identify false news through close scrutiny of a headline differs from the typical approach of professional fact checkers, who usually use "lateral reading" of alternative sources to corroborate claims (46).

We therefore conducted preregistered survey experiments in both the United States and India examining the effectiveness of presenting people with "tips" to help spot false news stories. [The US and India studies were each preregistered with Evidence in Governance and Politics; see *Materials and Methods*. All preregistered analyses are reported in this article or in the replication archive for the study (47).] Strikingly, our results indicate that exposure to variants of the Facebook media literacy intervention reduces people's belief in false headlines. These effects are not only an artifact of greater skepticism toward all information—although the perceived accuracy of mainstream news headlines slightly decreased, exposure to the intervention widened the gap in perceived accuracy between mainstream and false news headlines overall. In the United States, the effects of the treatment were particularly strong and remained statistically measurable after a delay of approximately 3 wk. These findings suggest that efforts to promote digital media literacy can improve people's ability to distinguish between false and mainstream news content, a result with important implications for both scientific research into why people believe misinformation online and policies designed to address the problem.

Our main research hypotheses evaluate whether the media literacy intervention reduces belief in false news stories (hypothesis 1 [H1]), increases belief in mainstream news content (H2), and improves respondents' ability to distinguish between them (H3). We also consider three research questions (RQs) for which our a priori expectations were less clear. First, past research shows that the effects of many experimental treatments (e.g., in persuasion and framing studies) decay quickly over time (48), although providing participants with novel information may have more long-lasting effects (49). We therefore test the durability of our treatment effect by leveraging a two-wave panel design to tests its effects several weeks after the initial intervention (RQ1). Second, it is also possible that interventions may work only to make individuals more skeptical of noncongenial content they are already inclined to dismiss, leaving their vulnerability to ideologically consistent misinformation unchanged. We therefore test for the heterogeneity of the treatment effects based on the partisan congeniality of the content (RQ2). Finally, we test whether the intervention changed self-reported intentions to share false stories or subsequent online news consumption behavior in the US sample where these measures were available (RQ3). Additional analyses exploring heterogenous treatment effects and alternate outcomes are discussed below, but full models appear in *SI Appendix*, section C. These analyses include whether intuitive cognitive style or prior headline exposure moderates the treatment effect, as well as whether the treatment affects the perceived credibility of "hyperpartisan" headlines.

Results

US Survey Experiment. Consistent with our first hypothesis (H1), randomized exposure to the media literacy intervention causes a decrease in the perceived accuracy of false news articles. Results from wave 1 of the US study in Table 1 show a decrease of nearly 0.2 points on a 4-point scale (intent to treat [ITT]: $\beta = -0.196$, $\mathrm{SE} = 0.020$; $P < 0.005$). We observe similar effects of the media literacy intervention on the perceived accuracy of hyperpartisan headlines (ITT: $\beta = -0.176$, $\mathrm{SE} = 0.020$; $P < 0.005$) (*SI Appendix*, section C, Table C2).

One concern is that the intent-to-treat effects described above understate the true effect of the intervention, which may have been neglected by some respondents. While we can offer the opportunity to read the digital literacy "fake news tips" intervention to a random subset of respondents, we cannot force every respondent to read these tips carefully.

We therefore also estimate the effect of the treatment on those who actually received it, which is known as the average treatment effect on the treated (ATT), using an instrumental variables approach. In this model, our indicator for receipt of treatment is the ability to correctly answer a series of follow-up questions about the content of the news tips (approximately two-thirds of respondents in the treatment condition [66%] were successfully treated) and our instrument is the original random assignment. Table 1 reports the ATT, which we compute using two-stage least-squares regression. With this approach, we estimate that the perceived accuracy of false headlines decreased by nearly 0.3 points on a 4-point scale (ATT: $\beta = -0.299$, $\mathrm{SE} = 0.030$; $P < 0.005$).[†]

We compare the characteristics of respondents who would successfully take the treatment only if assigned to it ("compliers") to those who would not even if assigned to treatment ("never takers") (*SI Appendix*, section B) (50). Compliers were more likely to be older, college graduates, interested in politics, politically knowledgeable, Republican identifiers, and more polarized in their feelings toward the two political parties than never takers. Compliers also scored lower in conspiracy predispositions and their feelings toward Donald Trump. However, the substantive magnitudes of most of these differences are modest (*SI Appendix*, section B, Fig. B1). Crucially, we find no statistically significant evidence that respondents who take the treatment differ in their baseline propensity to visit untrustworthy websites compared to those who do not (analysis conducted among participants for whom presurvey behavioral data are available; see *SI Appendix*, section A for details). The average number of prior visits to false news websites is actually greater among compliers than among never takers but this difference does not reach conventional thresholds of statistical significance (0.35 compared to 0.18; $P = 0.08$).

Our next hypotheses predicted that the media literacy intervention would increase the perceived accuracy of mainstream news (H2) and increase people's ability to successfully distinguish between mainstream and false news articles (H3). These results are shown in the second and third columns in Table 1. We find that exposure to the media literacy intervention had a small negative effect on belief in mainstream news in wave 1 (ITT, $\beta = -0.046$ [$\mathrm{SE} = 0.017$], $P < 0.01$; ATT, $\beta = -0.071$ [$\mathrm{SE} = 0.026$], $P < 0.01$). However, the negative effects of the intervention on the perceived accuracy of false news described above are larger. As a result, the media literacy intervention increased discernment between mainstream and false stories (ITT, $\beta = 0.146$ [$\mathrm{SE} = 0.024$], $P < 0.005$; ATT, $\beta = 0.223$ [$\mathrm{SE} = 0.035$], $P < 0.005$), demonstrating that it helped respondents to better distinguish between these two types of content. In relative terms, this effect represents a 26.5% improvement in respondents' ability to distinguish between mainstream and false news stories compared to the control condition.

Table 1. Effect of US media literacy intervention on perceived accuracy by news type

	False	Mainstream	Mainstream–false
ITT effects			
Media literacy intervention	−0.196***	−0.046**	0.146***
	(0.020)	(0.017)	(0.024)
Constant			0.551***
			(0.016)
Headline fixed effects	✓	✓	
N (headlines)	9,813	19,623	
N (respondents)	4,907	4,907	4,907
ATT			
Media literacy intervention	−0.299***	−0.071**	0.223***
	(0.030)	(0.026)	(0.035)
Constant			0.551***
			(0.016)
Headline fixed effects	✓	✓	
N (headlines)	9,813	19,623	
N (respondents)	4,907	4,907	4,907

*$P < 0.05$, **$P < 0.01$, ***$P < 0.005$ (two-sided). Data are from wave 1 (November to December 2018). Cell entries are ordinary least squares (OLS) or two-stage least-squares coefficients with robust standard errors in parentheses (clustered by respondent for false and mainstream news accuracy). Dependent variables for perceived false and mainstream news accuracy are measured on a 1 to 4 scale, where 1 represents "not at all accurate" and 4 represents "very accurate." The dependent variable for the difference in perceived false versus mainstream news accuracy is calculated at the respondent level as the mean difference in perceived accuracy between all false and all mainstream news headlines viewed.

In addition, we test the durability of these treatment effects in wave 2 per RQ1. After a delay between waves that averaged several weeks, the effect of the media literacy intervention on the perceived accuracy of false headlines remains statistically distinguishable from zero (*SI Appendix*, section C, Table C1). The median interval between waves was 20 d; the 5th to 95th percentile range was 16 to 29 d. While the effect is still present weeks later, its magnitude attenuates by more than half relative to wave 1 (ITT, $\beta = -0.080$ [$\mathrm{SE} = 0.019$], $P < 0.005$; ATT, $\beta = -0.121$ [$\mathrm{SE} = 0.028$], $P < 0.005$). In addition, the negative effect of the media literacy treatment on the perceived accuracy of mainstream news content was no longer statistically measurable by wave 2. As a result, the perceived accuracy difference between mainstream and false headlines remained statistically distinguishable from zero in the second wave, although its magnitude decayed ($\beta = 0.050$; $\mathrm{SE} = 0.020$; $P < 0.05$).

Fig. 1 illustrates the substantive magnitude of the intent to treat effects of the media literacy intervention in the United States using a binary indicator of perceived headline accuracy. The proportion of respondents rating a false headline as "very accurate" or "somewhat accurate" decreased from 32% in the control condition to 24% among respondents who were assigned to the media literacy intervention in wave 1, a decrease of 7 percentage points. This effect represents a relative decrease of

[†] These results were not preregistered but were estimated to match the preregistered compliance analyses reported in the India study. We also provide additional exploratory results that instead define compliance as answering each comprehension question correctly by the third try in *SI Appendix*, section C, Table C7. Our ATT effect estimates are necessarily smaller using this less stringent definition of treatment uptake.

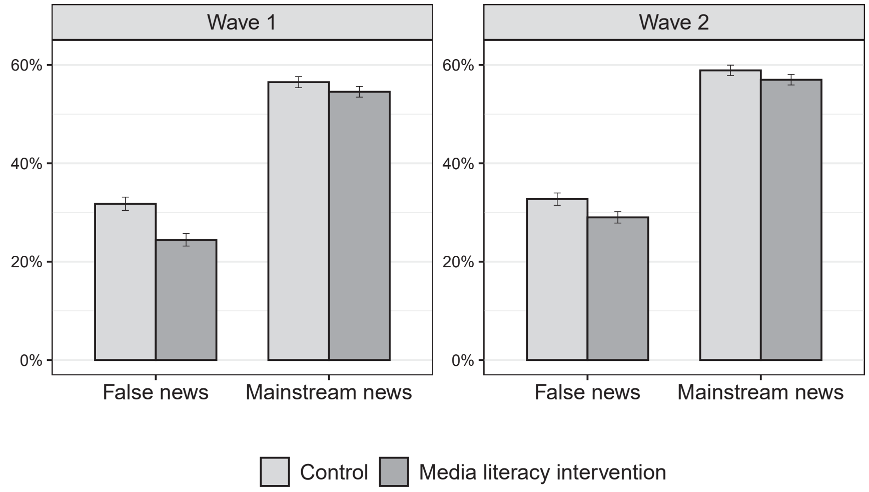

Fig. 1. Percentage of US respondents rating false and mainstream news headlines as somewhat accurate or very accurate. Respondents rated two and four headlines, respectively, in wave 1 and four and eight headlines, respectively, in wave 2. Headlines were selected randomly in wave 1, balanced by partisan congeniality, and presented in random order. Error bars are 95% confidence intervals of the mean.

approximately one-fourth in the percentage of people wrongly endorsing misinformation. Treatment effects continue to persist with this alternate measure—in wave 2, the intervention reduced the proportion of people endorsing false headlines as accurate from 33 to 29%, a 4-percentage-point effect. By contrast, the proportion of respondents who classified mainstream news as not very accurate or not at all accurate rather than somewhat or very accurate decreased only from 57 to 55% in wave 1 and 59 to 57% in wave 2.

Finally, RQ2 explores whether the effects of the media literacy intervention are moderated by the partisan congeniality of the headlines people rated. We find no consistent evidence that the effects of media literacy intervention are moderated by congeniality (*SI Appendix*, section C). In other words, the effects of the intervention were not differentially concentrated among headlines that were uncongenial to respondents—an encouraging null result that echoes findings in recent studies (34, 51, 52).

Additional results reported in *SI Appendix*, section C indicate that we have insufficient evidence to conclude that the intervention had an effect on self-reported intention to share false news or behavioral measures of news consumption (*SI Appendix*, section C, Table C13). However, the intervention did significantly increase sharing intentions for mainstream news and decrease sharing intentions for hyperpartisan news. This is consistent with previous studies that have reported mixed effects of warning labels on sharing intent (33, 34). The mixed results we observe for sharing intent may be attributable to the fact that belief accuracy questions appeared immediately before the sharing intent questions in the survey, which may prime accuracy concerns among respondents and thereby alter both real and self-reported sharing behavior. In addition, we find no measurable effect of the intervention on posttreatment visits to false news, mainstream news, or fact-checking sites, although these effects can be estimated only among the subset of respondents for whom we have behavioral data (*SI Appendix*, section C, Tables C14–15).

India Survey Experiments. As detailed in *Materials and Methods*, we conducted separate online and face-to-face surveys in India of different populations. For the online sample, we again find support for H1. The media literacy treatment significantly reduced beliefs in false news stories (ITT: $\beta = -0.126$, SE = 0.026; $P < 0.005$) in the first wave of a two-wave survey (Table 2). As with the US analysis, the ATT estimate was substantially larger (Table 2), although the difference between the ITT and the ATT is larger for the Indian analysis because fewer respondents answered all comprehension checks correctly in the Indian sample (28% in the online sample versus 66% in the United States).[‡] Respondents to the online survey who received the treatment were nearly half of a response category more skeptical of false news stories (ATT: $\beta = -0.470$, SE = 0.097; $P < 0.005$).

As in the United States, we also find no support for H2, which predicted that exposure to the treatment would increase the perceived accuracy of mainstream news. Instead, the perceived accuracy of mainstream news decreased, although by less than the effect observed for false news (ITT, $\beta = -0.071$ [SE = 0.025], $P < 0.01$; ATT, $\beta = -0.259$ [SE = 0.095], $P < 0.01$). Results again mirror the US study for H3—respondents better distinguished between mainstream and false articles (ITT, $\beta = 0.063$ [SE = 0.025], $P < 0.05$; ATT, $\beta = 0.221$ [SE = 0.088], $P < 0.05$).[§] While the magnitude of this effect is lower than for the US sample, it translates to a 17.3% improvement in discernment between mainstream and false news relative to the difference observed in the control condition. As we discuss in more detail below, treatment effects cannot be distinguished from zero in the second wave (RQ1).

While the analyses of the online samples from the United States and India show substantially similar results, results from the face-to-face survey in India differ in important ways. As shown in Table 2, we find no evidence that the treatment increased the perceived accuracy of mainstream news articles as predicted by H2. However, it did not reduce the perceived accuracy of these headlines either as we found in the United States and online studies. In addition, unlike the other studies, we find no evidence that the media literacy treatment systematically affected beliefs in false news stories (H1) or discrimination between false and mainstream news (H3) among India face-to-face respondents.

[‡] The analysis of compliers is presented in *SI Appendix*, section B. As *SI Appendix*, Fig. B2 shows, compliers in the India online sample (those who would take the treatment if assigned) were more likely to be young, male, Hindu, and high caste; to have graduated from college; to use WhatsApp; and to have more political knowledge and interest than never takers (respondents who would not take the treatment if assigned to receive it). We find no significant differences between these groups in the face-to-face sample (*SI Appendix*, section B, Fig. B3).

[§] In an exploratory analysis, we show that the result is robust to using an indicator for false news headlines instead of headline fixed effects (*SI Appendix*, section D, Table D3).

Table 2. Effect of India media literacy intervention on perceived accuracy by news type

	Online sample			Face-to-face sample		
	False news	Mainstream news	Mainstream−false	False news	Mainstream news	Mainstream−false
ITT effects						
Media literacy intervention	−0.126***	−0.071**	0.063*	−0.007	0.002	0.006
	(0.026)	(0.025)	(0.025)	(0.024)	(0.024)	(0.030)
Constant			0.361***			0.237***
			(0.017)			(0.021)
Headline fixed effects	✓	✓		✓	✓	
N (headlines)	17,031	17,163		13,712	13,969	
N (respondents)	3,177	3,182	3,160	3,267	3,314	3,140
ATT						
Media literacy intervention	−0.470***	−0.259**	0.221*	−0.035	0.011	0.028
	(0.097)	(0.095)	(0.088)	(0.113)	(0.113)	(0.138)
Constant			0.361***			0.237***
			(0.017)			(0.021)
Headline fixed effects	✓	✓		✓	✓	
N (headlines)	17,031	17,163		13,712	13,969	
N (respondents)	3,177	3,182	3,160	3,267	3,314	3,140

*$P < 0.05$, **$P < 0.01$, ***$P < 0.005$ (two-sided). Data are from wave 1 (April to May 2019). Cell entries are OLS or two-stage least-squares coefficients with robust standard errors in parentheses (clustered by respondent for false and mainstream news accuracy). Dependent variables for perceived false and mainstream news accuracy are measured on a 1 to 4 scale where 1 represents "Not at all accurate" and 4 represents "Very accurate." The dependent variable for the difference in perceived false versus mainstream news accuracy is calculated at the respondent level as the mean difference in perceived accuracy between all false and all mainstream news headlines viewed.

We directly assess the difference between the two samples using a pooled model. These results, presented in *SI Appendix*, section D, Table D13, indicate that we can reject the null of no difference in media literacy intervention effects between the face-to-face and online samples for ratings of false news (ITT estimate of the face-to-face − online difference, $\beta = 0.118$ [SE = 0.035], $P < 0.005$; ATT, $\beta = 0.428$ [SE = 0.149], $P < 0.005$) and for the ITT estimate for ratings of mainstream news stories (ITT, $\beta = 0.072$ [SE $= 0.035$], $P < 0.05$; ATT, $\beta = 0.266$ [SE $= 0.148$], $P > 0.05$), but not for the average difference in perceived accuracy between mainstream and false news stories (ITT, $\beta = -0.057$ [SE = 0.039], $P > 0.05$; ATT, $\beta = -0.194$ [SE = 0.163], $P > 0.05$). Potentially, the differences in our results between studies may reflect the different survey modes or demographic compositions of the samples (we consider this issue further in *Discussion*).

Fig. 2 illustrates the substantive magnitudes of the ITT effect for respondents to the two India surveys. For the online survey, exposure to the media literacy intervention reduced the percentage of respondents rating false headlines as somewhat accurate or very accurate from 49% in the control group to 44% in the treatment group, a decrease of approximately 10% in relative terms. As noted above, the effect of the intervention was much greater for those who received the treatment successfully—the ATT estimate indicates a decline of approximately 19 percentage points in endorsement of false headlines among this group (*SI Appendix*, section D, Table D7). By contrast, although mainstream stories were also viewed more skeptically by online survey respondents who received the media literacy intervention (from 63% for controls to 60% in the treatment group), the relative decrease in perceived accuracy was only half of what was observed for false headlines. Finally, as noted above, there was no significant difference on average between face-to-face survey respondents who received the media literacy treatment and those who did not for either false or mainstream headlines (belief levels were higher overall among face-to-face respondents—62% for false headlines and 72% for mainstream headlines).

A key research question was whether any treatment effects would persist over time (RQ1). We found no statistically reliable evidence that the treatment affected headline accuracy ratings among wave 2 respondents in either India sample (*SI Appendix*, section D, Table D2).[¶] Finally, we did not find statistically reliable evidence in either India survey that the media literacy intervention's effects were moderated by partisan congeniality (RQ2; *SI Appendix*, section D, Table D9). We cannot conclude that the effects we observe depend on whether the headlines were congenial to respondent partisanship.

Discussion

Comparing our results across studies reveals a relatively consistent pattern. As Fig. 3 indicates, both the US study and the India online study find negative effects on the perceived accuracy of false headlines that are diminished for mainstream news headlines. As a result, respondents' ability to discern between mainstream and false news increased. Although these findings are not observed in the India face-to-face study, a combined estimate pooling data from all three studies replicates the overall pattern of reduced false news accuracy perceptions and increased discernment between mainstream and false news. The treatment effect estimates for each study as well as the pooled results are shown in Fig. 3 (see *SI Appendix*, section F for full model results).

These effects are also substantively meaningful. Although our study does not instruct respondents to apply the lessons from the intervention to the headline ratings task, the effect sizes are comparable to the estimated effect of exposure to the labels Facebook initially used to indicate that articles were disputed by fact checkers (ITT estimate: $\beta = -0.236$, SE $= 0.036$; $P < 0.005$) (34) and greatly exceed the effects of a general warning about false news in the same study (ITT estimate: $\beta = -0.079$, SE $= 0.034$; $P < 0.05$). A comparison of effect sizes with other

[¶] The median intervals between waves were 20 d for the face-to-face survey and 19 d in the online survey; the 5th to 95th percentile ranges were 14 to 29 d and 15 to 26 d, respectively. The wave 2 results are substantively unchanged if we include the four additional false headlines from the fact-check message experiment described in *SI Appendix*, section A (*SI Appendix*, section D, Table D10).

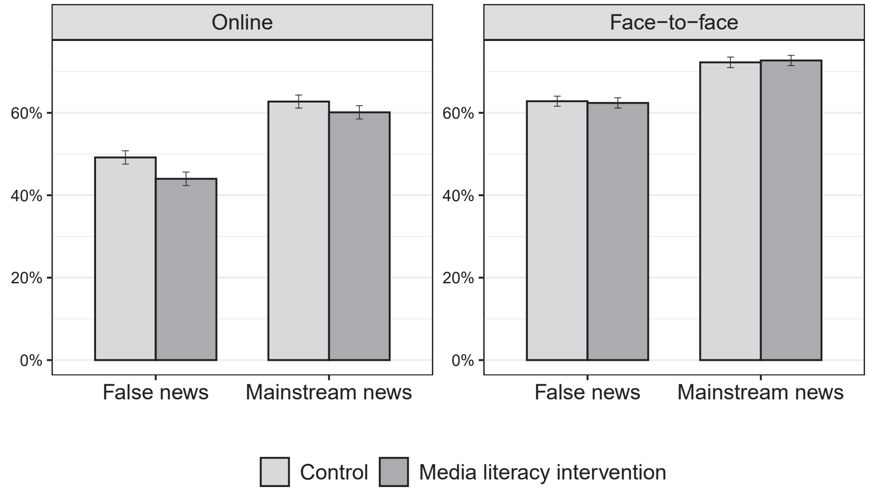

Fig. 2. Percentage of India respondents rating false and mainstream news headlines as somewhat accurate or very accurate in wave 1. Respondents rated six of each type of headline. The headlines were balanced by partisan congeniality and presented in random order. Error bars are 95% confidence intervals of the mean.

randomized media literacy interventions in *SI Appendix*, section E, Table E1 shows that our US study has the largest measured effect size to date on ratings of false headlines ($d = 0.20$) and that the India online study ($d = 0.11$) falls in the upper middle of the distribution. Moreover, effect sizes are substantially larger for respondents who were successfully treated with the media literacy intervention.

Despite the strength of the combined estimates, the effectiveness of the intervention varied across samples. First, the intervention may have been more unfamiliar or difficult to understand for Indian respondents, who successfully received the treatment at a much lower rate than those in the United States. Additional evidence suggests that respondents may have applied the intervention differently. Specifically, the US data show that the negative effects of the media literacy intervention on perceived headline accuracy were greater for headlines from untrustworthy, hyperpartisan, and unfamiliar mainstream sources that respondents in the control group found less plausible to begin with ($r = 0.79$; *SI Appendix*, section C, Table C11 and Fig. C2). This finding helps explain the observed negative effect of the media literacy intervention on the perceived accuracy of mainstream news overall.[#] By contrast, no such relationship between baseline headline accuracy and media literacy intervention effects is observed in the Indian online data ($r = -0.09$; *SI Appendix*, section D8), suggesting respondents became more skeptical across the board.

Problems applying the intervention may have been particularly acute for respondents in the face-to-face sample. This group is quite dissimilar from both the highly educated online sample in India and the US sample on a number of important indicators (*SI Appendix*, section B). In particular, participants in the face-to-face study had much less experience with the task of evaluating news headlines online—only 11% reported using WhatsApp compared with 90% for the online sample in India. Correspondingly, an exploratory analysis shows the effects of the intervention were similar among WhatsApp users across samples; however, these effects are imprecisely estimated among face-to-face participants because WhatsApp use was so rare (*SI Appendix*, section D9).

Finally, we consider the potential trade-off between increased skepticism toward false news headlines and decreased belief in mainstream news headlines. Our results do indicate that increased skepticism of false news headlines may come at the expense of decreased belief in mainstream news headlines—the media literacy intervention reduced the perceived accuracy of these headlines in both the US and India online surveys. However, the magnitude of the decrease in the perceived accuracy of mainstream news headlines ranges from under 25% (United States) to just over half (India online sample) of the estimated size of the decrease in the perceived accuracy of untrustworthy news headlines in wave 1 of our surveys. As a result, respondents' overall ability to distinguish between mainstream and untrustworthy news increases by more than 26% in the US sample and 17% in the highly educated online Indian sample. Moreover, we observe no measurable decrease in the perceived accuracy of mainstream news headlines in wave 2 of any of our surveys.

A related concern is that the intervention could reduce the overall accuracy of people's beliefs given that they typically consume much more information from mainstream sources than from untrustworthy ones (53). To address this concern, we use US Pulse web metering data to estimate the overall change that the intervention would hypothetically induce in people's ability to accurately discern credible news given exposure rates for US participants to different types of news sources (see *SI Appendix*, section G for details). Because Americans' news consumption is concentrated among the high-prominence mainstream outlets for which the intervention may have had a small positive effect[#], these calculations indicate that individuals would reach valid accuracy beliefs for 64.6% of stories in the treatment group compared to 62.9% in the control group. Moreover, the percentage of "false positives"—stories they encounter from dubious sources and believe to be true—would decrease from 6.1% of all stories consumed in the control group to 4.9% in the treatment group.

Conclusion

The findings we present provide important evidence that shortfalls in digital media literacy are an important factor in why people believe misinformation that they encounter online. We find that a simple, scalable media literacy intervention can decrease the perceived accuracy of false news content and help

[#] An exploratory analysis of whether source prominence moderates the effects of the media literacy intervention shows that the negative effects we observe for the perceived accuracy of mainstream news headlines were concentrated among stories from low-prominence sources. By contrast, we find that the intervention appeared to increase the perceived accuracy of stories from high-prominence sources (*SI Appendix*, section C9).

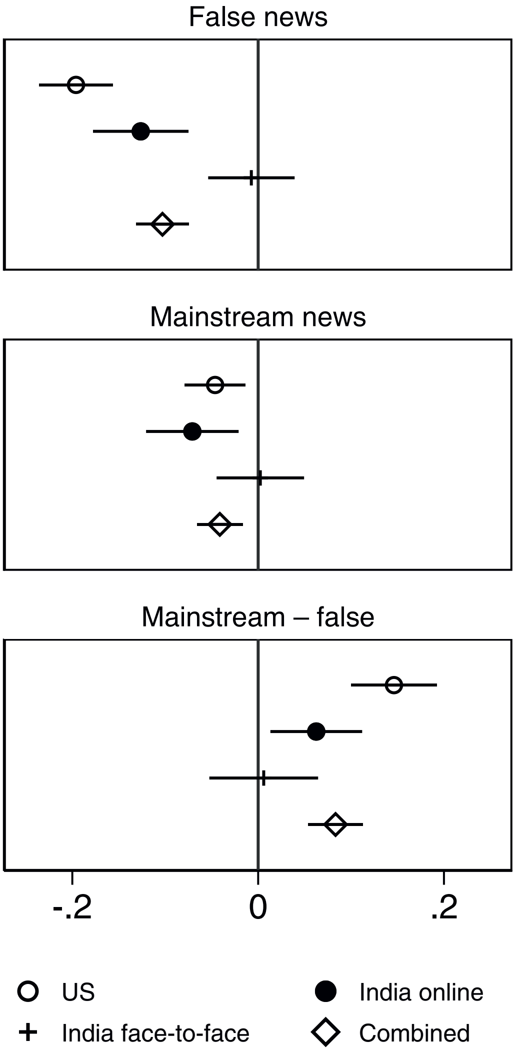

Fig. 3. Data are from wave 1. Effect sizes are plotted with 95% confidence intervals. Effect sizes are estimated at the headline level for false and mainstream news and at the respondent level for the difference in perceived accuracy between them.

people to better distinguish it from factual mainstream news in both the United States and India. Moreover, the improvement in headline accuracy rating performance we observe does not depend on whether the claims in question align with respondents' political predispositions.

Our results further suggest that media literacy campaigns could be an effective strategy to help counter false or misleading news, a finding with important real-world implications. Some explanations for belief in misinformation identify factors that resist intervention (4), while others propose policy approaches that are effective in practice but difficult to scale (34, 44). Inoculation, while effective in preemptively refuting misinformation in specific domains (27–30), may not improve discernment when evaluating a diverse array of real-world news content. By contrast, these results show that a brief intervention which could be inexpensively disseminated at scale can be effective at reducing the perceived accuracy of false news stories, helping users more accurately gauge the credibility of news content they encounter on different topics or issues.

Although these results suggest that digital literacy interventions could be a valuable tool in the fight against misinformation, several caveats should be offered. First, the effect sizes were modest; this simple intervention did not eliminate belief in false news headlines. Second, the effects decayed over time (diminishing in magnitude in the United States and no longer remaining statistically measurable in the India online study), suggesting the need for social media companies, journalists, and educators to reinforce these lessons on a recurring basis. Third, although the intervention improved overall discernment between mainstream and false news headlines, it did have a small but measurable negative effect on the perceived accuracy of mainstream news stories. Fourth, all treated participants were exposed to the intervention; many real-world Facebook users are likely to have ignored the Tips to Spot False News link when the company offered it in their feeds. (The difference between our ITT and ATT estimates illustrates how lack of attention to the treatment reduces its effectiveness.) Finally, we have insufficient evidence to conclude that the intervention changed real-world consumption of false news, perhaps because information habits are ingrained and difficult to alter. However, we do find evidence that the treatment increased respondents' intent to share mainstream news and decreased intent to share hyperpartisan news, suggesting the possibility of changes to social media behavior.

Our findings also suggest a number of directions for future research. One advantage of the study is that we used an actual intervention deployed globally by a technology company that has served as an important access point to false news (53). However, scholars should conduct comparative evaluations of the effects of different interventions rather than relying on this model as a default and test the effectiveness of these approaches in different samples, countries, and electoral contexts. These evaluations should include tests of more intensive digital literacy training models (such as the "lateral reading" approach used by professional fact checkers), which could potentially have larger and/or more durable effects. In addition, scholars should seek to better understand the mechanism through which such interventions operate, identifying whether the effects they observe are due to increased accuracy concerns versus helping people learn more effective heuristics for evaluating news content. Finally, researchers should further examine whether and how media literacy interventions can increase the frequency or effectiveness of accuracy-promoting behavior in social contexts. Even if these interventions do not reach everyone, improving the media literacy of a subset of the population could yield wider positive effects if, for instance, those who are treated help to correct the misinformation they see on social media (54).

To our knowledge, though, these findings are the most compelling demonstration to date that a real-world digital literacy intervention can have significant and potentially lasting effects. While efforts to improve online digital literacy are not a panacea, they may prove to be a vital and cost-effective way to reduce people's vulnerability to false news and thereby improve the information health of democracies.

Materials and Methods

Data Collection. We conducted two-wave panel surveys of respondents that included an embedded media literacy intervention. One survey took place in the United States and two were conducted in India. All took place during periods of high political interest during and immediately after national electoral campaigns.

In the United States, we conducted a two-wave online panel survey fielded by the survey company YouGov shortly after the 2018 US midterm elections (wave 1, November 20 to December 27, 2018, $N = 4,907$; wave 2, December 14, 2018 to January 3, 2019, $N = 4,283$).[‖] Respondents were selected by YouGov's matching and weighting algorithm to approximate the demographic and political attributes of the US population (32%

[‖] Respondents took the surveys in order with a delay between them averaging approximately 3 wk. The dates above reflect the ranges in which all wave 1 and wave 2 responses were collected.

college graduates, 45% male, median age 50 y; 46% identify as Democrats and 36% as Republicans). A subset of these respondents were members of the YouGov Pulse panel and voluntarily provided behavioral data on their online information consumption as well (see *SI Appendix*, section A for more details).

US data collection was approved by the Institutional Review Boards (IRBs) at the University of Michigan (HUM00153414), Washington University in St. Louis (201806142), and Princeton University (10875). University of Exeter accepted the University of Michigan IRB approval. All subjects gave consent to participate in our study. The US study preanalysis plan is available at https://osf.io/u3sgc.

For India, we conducted two separate two-wave panel studies, one online and the other face to face. Both surveys were conducted in Hindi. Respondents were excluded if indicated they mostly or always give humorous or insincere answers to survey questions (which amounted to 7–8% of responses in the online survey by wave compared to less than 1% in the face-to-face survey; this exclusion represents a deviation from our preregistration, but the results in Table 2 are robust to including these respondents). In the online survey, we collected survey data from a national convenience sample of Hindi-speaking Indians recruited via Mechanical Turk and the Internet Research Bureau's Online Bureau survey panels (wave 1, April 17 to May 1, 2019, $N = 3,273$; wave 2, May 13 to 19, 2019, $N = 1,369$). The India face-to-face survey was conducted by the polling firm Morsel in Barabanki, Bahraich, Domariyaganj, and Shrawasti, four parliamentary constituencies in the state of Uttar Pradesh where Hindi is the dominant language (wave 1, April 13 to May 2, 2019, $N = 3,744$; wave 2, May 7 to 19, 2019, $N =$ 2,695). These locations were chosen, in part, due to their higher levels of religious polarization, which we anticipated might increase demand for and belief in online misinformation and rumors. The representative random sample for the India face-to-face survey was drawn from the public voter registration list for these constituencies and was administered orally by trained enumerators to account for low literacy rates. Relative to the face-to-face survey, online respondents were more likely to be male (72% versus 64%), younger (median age 30 y versus 37 y), more educated (76% college graduates versus 6%), higher caste (42% low caste versus 74% in the face-to-face sample), more active on social media (90% use WhatsApp versus 11%), more interested in politics (3.7 versus 2.9 on a 4-point scale), more knowledgeable about politics (providing correct responses to an average of 2.8 vs. 2.1 of four true–false questions about Indian politics), and slightly less likely to support the BJP (42% versus 46%) (*SI Appendix*, section D, Table D1).

India data collection was approved by the IRBs at the University of Michigan (HUM00160358), Ashoka University, and Morsel Research and Development (HIRB000007598). University of Exeter accepted the University of Michigan IRB approval. All subjects gave consent to participate in our study. The India study preanalysis plan is available at https://osf.io/97rnz.

Our study contexts can be viewed as a most-different case comparison among democracies (55). India and the United States are broadly considered the poorest and richest countries, respectively, in terms of income per capita among longstanding large democracies (ref. 56, p. 42). As a result, India is likely to have lower levels of education and media literacy than the United States, which raises questions about the efficacy of any media intervention. The two studies we conduct within India further refine this comparison, allowing us to evaluate the effects of the media literacy intervention among both an online sample that has demographics that are more similar to the United States and a face-to-face sample in one of the poorest regions in India. Our study can thus provide evidence about the efficacy of a media literacy intervention across democratic contexts that differ in levels of income, education, and digital media use.

News Headline Rating Task. The main outcome of interest in all three surveys was the perceived accuracy of mainstream and false news headlines. To construct this measure, we asked respondents to evaluate the accuracy of a number of headlines on a 4-point scale ranging from very accurate (4) to not at all accurate (1). All of the headlines were published by actual news sources or circulated on Facebook or WhatsApp within 6 mo of the respective survey, and a portion of the headlines were rated as false by at least one third-party fact-checking organization. The order of the headlines was randomized within wave for each respondent. All headlines are shown in *SI Appendix*, section H1.

In the US survey, respondents evaluated 16 different headlines that varied across multiple dimensions: news type (mainstream versus hyperpartisan versus false), valence (pro-Democrat versus pro-Republican), and prominence among mainstream sources (high versus low). We define high-prominence mainstream sources as those that more than 4 in 10 Americans recognize in Pew polling (57). Hyperpartisan stories are those that are technically factual but present slanted facts in a deceptive manner. We selected these stories from hyperpartisan sources identified in prior work (58) (*SI Appendix*, section H). This process resulted in 8 mainstream, 4 false, and 4 hyperpartisan headlines. In wave 1, respondents were shown 8 headlines (a randomly selected headline from the two available for each possible combination of news type, valence, and source prominence), while respondents in wave 2 were shown all 16 headlines. Headlines were presented as they would appear on the Facebook news feed to replicate a typical decision environment. Specifically, respondents were shown the article previews that are automatically generated by Facebook when a link is entered into the news feed that featured a headline, a photo, the news source's web domain, and in certain cases a byline or story text snippet. Respondents were asked to rate the accuracy of each headline.

In the India surveys, we adopted the same approach in asking respondents to evaluate the accuracy of headlines that varied across several dimensions: valence (congenial to BJP supporters versus congenial to BJP opponents) and accuracy (true articles from mainstream sources[††] versus false articles as identified by fact checkers). Nationalism is also commonly linked to misinformation in India (59). The issue was particularly salient when the India surveys were conducted (a time of escalating tensions between India and Pakistan), so we also asked respondents to rate the accuracy of true and false headlines relevant to nationalist concerns in the country (either India–Pakistan or Hindu–Muslim relations). Unlike the US study (where the same headlines were used in both waves 1 and 2 to test for prior exposure effects), we used different sets of headlines in each wave. Finally, 4 additional false headlines were included in the second wave based on fact checks conducted between the two waves.[‡‡] In total, respondents rated 12 headlines in wave 1 (6 false and 6 true) and 16 in wave 2 (10 false and 6 true). Respondents were presented with the headline in text format in the online survey, while enumerators read the headlines to respondents in the face-to-face survey. In both cases, participants were asked to evaluate the accuracy of all headlines they were presented in each wave.

Analysis. Our primary analyses are pooled OLS models predicting perceptions of headline accuracy on a four-point scale that ranges from not at all accurate to very accurate. These models were estimated at the headline level with fixed effects for each headline. Although we attempted to choose stories that were balanced in their face validity, the headlines differed in plausibility because the actual stories were not constructed by researchers. We therefore use the fixed effects to account for baseline differences in perceived accuracy between headlines. Because respondents rated multiple headlines, we also compute cluster-robust standard errors. In addition to the pooled OLS models, we also examine the difference in accuracy beliefs between mainstream and false headlines at the respondent level by calculating a respondent-level measure of the difference in mean levels of perceived accuracy between mainstream and false headlines. Higher scores on this scale indicate better ability to discern between stories of different types.

Congeniality is a binary variable that is coded at the headline level for partisans to indicate whether a story is consistent with a respondent's partisan leanings (e.g., a Democrat evaluating a story that is favorable to a Democrat would take the value of 1). Uncongenial is coded as the opposite. The baseline category is reserved for headline ratings by pure nonpartisans. To determine the partisanship of respondents in the US survey, we used the standard two-question party identification battery (which includes leaners) to classify respondents as Democrats or Republicans. Because India has a multiparty system, we classified respondents there as BJP supporters if they reported liking the BJP more than any other party (on a four-point scale) and as a BJP opponent if they liked any other party more than the BJP.

The key explanatory variable of interest is exposure to the media literacy intervention, which was adapted from an intervention deployed by Facebook and WhatsApp around the world, including in Hindi-language newspapers in India (see *SI Appendix*, section A for details and the exact text). We randomly assigned respondents in wave 1 of the US and India studies with probability 0.5 to be exposed to a set of tips for distinguishing false news stories from mainstream stories. In the US survey experiment, 10

[††] Mainstream news sources included *ZeeNews*, *Washington Post*, *National Herald India*, *IndiaToday*, *Nikkei Asian Review*, Reuters, and Bloomberg.

[‡‡] These additional headlines were part of a parallel study; further details are provided in *SI Appendix*, section A).

tips published by Facebook were presented verbatim in 2 groups of 3 and one group of 4. In the India surveys, 6 tips from those published by Facebook and WhatsApp were presented in 2 groups of 3 after being adapted for the face-to-face format (omitting cues such as URLs that would not be present and reducing their length when possible). The treatment was administered before the headline rating task and respondents were asked comprehension questions after each group of tips to determine receipt of treatment. We calculate both ITT estimates using the full sample and the ATT below.

For calculating the ATT, receipt of treatment is defined as answering all comprehension questions correctly on the first try (online participants had up to three chances to answer correctly; face-to-face respondents had one chance). Receipt of treatment was substantially higher in the United States (66% in the US online survey and 24% and 28%, respectively, in the India face-to-face and online surveys). Our ATT estimates likely understate effects for two reasons. A small fraction of respondents may be misclassified as compliers because they guessed correctly on all of the comprehension checks, which will diminish our ATT estimates relative to the true effect (although the likelihood of such an outcome under random guessing is low relative to the estimated compliance rates). Additionally, some respondents we classify as receiving treatment are effectively "always takers"—people who either saw the tips in real life or have already internalized them through frequent experience. The intervention should have no effect on them. As such, our ATT estimates will understate the true effect.

Data Availability. Data files and scripts necessary to replicate the results in this article are available at the Dataverse repository: https://doi.org/10.7910/DVN/Q5QINN.

ACKNOWLEDGMENTS. We thank Democracy Fund, the European Research Council under the European Union's Horizon 2020 research and innovation program (Grant 682758), and the Facebook Integrity Foundational Research Awards for funding support and Sumit Parashar for outstanding research assistance. We are also grateful to Calvin Lai, Ben Kaiser, Ro'ee Levy, Jonathan Mayer, Kevin Munger, Yusuf Neggers, Christine Stedtnitz, Emily Thorson, Tim Vercellotti, Shun Yamaya, Dannagal Young, and Thomas Zeitzoff; to Henrik Oscarsson, Jesper Stromback, and the "Knowledge Resistance" group at the University of Gothenburg; and to seminar participants at Brigham Young University, George Washington University, Harvard Kennedy School, Katholieke Universiteit Leuven, and Stanford University for helpful comments. We thank YouGov, Morsel Research and Development, and the Internet Research Bureau for assistance administering the surveys. The India study was supported by funding from Facebook, but the company and its employees played no role in how the study was designed, conducted, or analyzed. All conclusions and any errors are our own.

1. C. Jackson, Fake news, filter bubbles, post-truth and trust: A study across 27 countries. Ipsos, 5 September 2018. https://www.ipsos.com/en-us/news-polls/Fake-News-Study. Accessed 14 August 2019.
2. D. Funke, D. Flamini, A guide to anti-misinformation actions around the world. Poynter Institute (2019). https://www.poynter.org/ifcn/anti-misinformation-actions/. Accessed 14 August 2019.
3. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 1–28 (2017).
4. G. Pennycook, D. G. Rand, Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2018).
5. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 US presidential election. *Science* **363**, 374–378 (2019).
6. Y. Eshet, Digital literacy: A conceptual framework for survival skills in the digital era. *J. Educ. Multimedia Hypermedia* **13**, 93–106 (2004).
7. P. Gilster, *Digital Literacy* (Wiley Computer Pub, New York, NY, 1997).
8. J. B. Horrigan, Digital readiness gaps. Pew Research Center, October 2019. https://www.pewresearch.org/internet/2016/09/20/digital-readiness-gaps/. Accessed 19 November 2019.
9. J. Poushter, C. Bishop, H. Chwe, Social media use continues to rise in developing countries but plateaus across developed ones. Pew Research Center, 19 June 2018. https://www.pewresearch.org/global/2018/06/19/social-media-use-continues-to-rise-in-developing-countries-but-plateaus-across-developed-ones/. Accessed 20 November 2019.
10. A. Perrin, M. Anderson, Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018. Pew Research Center, 10 April 2019. https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/. Accessed 20 November 2019.
11. C. Silverman, This analysis shows how fake election news stories outperformed real news on facebook. Buzzfeed, 16 November 2016. https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.ohXvLeDzK#.cwwgb7EX0. Accessed 16 December 2016.
12. M. Barthel, A. Mitchell, J. Holcomb, Many Americans believe fake news is sowing confusion. Pew Research Center (2016). https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion. Accessed: 27 May 2020.
13. T. McLaughlin, How WhatsApp fuels fake news and violence in India. Wired, 12 December 2018. https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/. Accessed 9 October 2019.
14. A. Thakar, India's fake-news crisis has intensified during the 2019 elections, say fact-checkers. Quartz India, 3 August 2019. https://qz.com/india/1609763/alt-news-boom-live-on-fake-news-detection-amid-indian-election/. Accessed 30 July 2019.
15. S. Poonam, S. Bansal, Misinformation is endangering India's election. The Atlantic, 1 April 2019. https://www.theatlantic.com/international/archive/2019/04/india-misinformation-election-fake-news/586123/. Accessed 30 July 2019.
16. V. Narayanan *et al.*, "News and information over Facebook and WhatsApp during the Indian election campaign" (Data Memo 2019.2, Project on Computational Propaganda, Oxford, UK, 2019).
17. S. Craft, S. Ashley, A. Maksl, News media literacy and conspiracy theory endorsement. *Communication Public* **2**, 388–401 (2017).
18. M. J. Abramowitz, Stop the manipulation of democracy online. New York Times, 11 December 2017. https://www.nytimes.com/2017/12/11/opinion/fake-news-russia-kenya.html. Accessed 30 July 2019.
19. T. C. Helmus *et al.*, *Russian Social Media Influence: Understanding Russian Propaganda in Eastern Europe* (Rand Corporation, 2018).
20. S. Patil, India has a public health crisis. It's called fake news. New York Times, 29 April 2019. https://www.nytimes.com/2019/04/29/opinion/india-elections-disinformation.html. Accessed 30 July 2019.
21. K. Jazynka, Colleges turn 'fake news' epidemic into a teachable moment. Washington Post, 6 April 2017. https://www.washingtonpost.com/lifestyle/magazine/colleges-turn-fake-news-epidemic-into-a-teachable-moment/2017/04/04/04114436-fd30-11e6-99b4-9e613afeb09f_story.html?utm_term=.46c23796b30f. Accessed 30 July 2019.
22. E. Conley-Keck, Illinois students could soon get lessons in fake news. WQAD, 3 March 2019. https://wqad.com/2019/03/03/illinois-students-could-soon-get-lessons-in-fake-news/. Accessed 30 July 2019.
23. WhatsApp and NASSCOM collaborate to teach about fake news. India Today, March 19 March 2019. https://www.indiatoday.in/education-today/news/story/whatsapp-and-nasscom-collaborate-to-teach-about-fake-news-1481882-2019-03-19. Accessed 30 July 2019.
24. E. K. Vraga, M. Tully, News literacy, social media behaviors, and skepticism toward information on social media. *Inf. Commun. Soc.*, 10.1080/1369118X.2019.1637445 (2019).
25. S. M. Jones-Jang, T. Mortensen, J. Liu, Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *Am. Behav. Sci.*, 10.1177/0002764219869406 (2019).
26. A. Huguet, J. Kavanagh, G. Baker, M. S. Blumenthal, *Exploring Media Literacy Education as a Tool for Mitigating Truth Decay* (RAND Corporation, 2019).
27. J. A. Banas, G. Miller, Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Hum. Commun. Res.* **39**, 184–207 (2013).
28. K. Braddock, Vaccinating against hate: Using attitudinal inoculation to confer resistance to persuasion by extremist propaganda. *Terrorism Polit. Violence*, 1–23 (2019).
29. J. Cook, S. Lewandowsky, U. K. Ecker, Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS One* **12**, e0175799 (2017).
30. A. Leiserowitz, S. Rosenthal, E. Maibach, E. Maibach, Inoculating the public against misinformation about climate change. *Global Chall.*, 10.1080/09546553.2019.1693370 (2017).
31. S. van der Linden, E. Maibach, J. Cook, A. Leiserowitz, S. Lewandowsky, Inoculating against misinformation. *Science* **358**, 1141–1142 (2017).
32. J. Roozenbeek, S. van der Linden, Fake news game confers psychological resistance against online misinformation. *Palgrave Commun.* **5**, 65 (2019).
33. G. Pennycook, A. Bear, E. T. Collins, D. G. Rand, The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Manag. Sci.*, 10.1287/mnsc.2019.3478 (2020).
34. K. Clayton *et al.*, Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Polit. Behav.*, 10.1007/s11109-019-09533-0 (2019).
35. A. Mosseri, A new educational tool against misinformation. Facebook, 6 April 2017. https://newsroom.fb.com/news/2017/04/a-new-educational-tool-against-misinformation/. Accessed 23 May 2017.
36. J. Constine, Facebook puts link to 10 tips for spotting 'false news' atop feed. TechCrunch, 6 April 2017. https://techcrunch.com/2017/04/06/facebook-puts-link-to-10-tips-for-spotting-false-news-atop-feed/. Accessed 19 March 2019.
37. G. Mezzofiore, Facebook fights fake news with print newspaper ads. Mashable, 8 May 2017. https://mashable.com/2017/05/08/facebook-fake-news-newspaper-ad-elections-uk/#PZqzuqCrwqqP. Accessed 19 March 2019.
38. T. Srivastav, Facebook turns to newspaper ads to combat fake news in India. The Drum, 22 September 2017. https://www.thedrum.com/news/2017/09/22/facebook-turns-newspaper-ads-combat-fake-news-india. Accessed 19 March 2019.
39. H. Tsukayama, Facebook fights fake news online with full-page print newspaper ads. Washington Post, 14 April 2017. https://www.washingtonpost.com/news/the-switch/wp/2017/04/14/facebook-fights-fake-news-online-with-full-page-print-newspaper-ads/?utm_term=.c736f9621a30. Accessed 19 March 2019.

40. A. Al-Heeti, Facebook will fight fake news with real newspaper ads (and more). CNet, 23 May 2018. https://www.cnet.com/news/facebook-is-fighting-misinformation-with-news-literacy-campaign-help-from-researchers/. Accessed 19 March 2019.
41. The Quint, After lynchings in India, 10 tips from WhatsApp to spot fake news. 10 July 2018. https://www.thequint.com/news/india/after-lynchings-in-india-whatsapp-offers-tips-to-spot-fake-news. Accessed 20 August 2019.
42. Agence France Presse, WhatsApp running anti-fake news ads in Pakistan ahead of elections. 19 July 2018. https://www.rappler.com/technology/news/207674-whatsapp-anti-fake-news-campaign-pakistan-elections. Accessed 20 August 2019.
43. E. W. Austin *et al.*, The effects of increased cognitive involvement on college students' interpretations of magazine advertisements for alcohol. *Commun. Res.* **29**, 155–179 (2002).
44. G. Pennycook, T. D. Cannon, D. G. Rand, Prior exposure increases perceived accuracy of fake news. *J. Exp. Psychol. Gen.* **147**, 1865–1880 (2018).
45. L. K. Fazio, D. G. Rand, G. Pennycook, Repetition increases perceived truth equally for plausible and implausible statements. *Psychon. Bull. Rev.* **26**, 1705–1710 (2019).
46. S. Wineburg, S. McGrew, Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teach. Coll. Rec.* **121**, 1–40 (2019).
47. A. Guess *et al.*, Data from "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." Harvard Dataverse. https://doi.org/10.7910/DVN/Q5QINN. Deposited 29 May 2020.
48. A. S. Gerber, J. G. Gimpel, D. P. Green, D. R. Shaw, How large and long-lasting are the persuasive effects of televised campaign ads? Results from a randomized field experiment. *Am. Polit. Sci. Rev.* **105**, 135–150 (2011).
49. A. Coppock, E. Ekins, D. Kirby, The long-lasting effects of newspaper op-eds on public opinion. *Q. J. Polit. Sci.* **13**, 59–87 (2018).
50. M. Marbach, D. Hangartner, Profiling compliers and non-compliers for instrumental variable analysis. *Polit. Anal.*, 10.1017/pan.2019.48 (2020).
51. B. Swire-Thompson, U. K. Ecker, S. Lewandowsky, A. J. Berinsky, They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Polit. Psychol.* **41**, 21–34 (2020).
52. M. J. Aird, U. K. Ecker, B. Swire, A. J. Berinsky, S. Lewandowsky, Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *R. Soc. Open Sci.* **5**, 180593 (2018).
53. A. Guess, B. Nyhan, J. Reifler, Exposure to untrustworthy websites in the 2016 U.S. election. *Nat. Hum. Behav.* **4**, 472–480 (2020).
54. L. Bode, E. K. Vraga, See something, say something: Correction of global health misinformation on social media. *Health Commun.* **33**, 1131–1140 (2018).
55. J. Seawright, J. Gerring, Case selection techniques in case study research: A menu of qualitative and quantitative options. *Polit. Res. Q.* **61**, 294–308 (2008).
56. A. Stepan, J. J. Linz, Y. Yadav, *Crafting State-Nations: India and Other Multinational Democracies* (Johns Hopkins University Press, 2011).
57. A. Mitchell, J. Gottfried, J. Kiley, K. E. Matsa, Political polarization & media habits. Pew Research Center, 21 October 2014. https://www.pewresearch.org/wp-content/uploads/sites/8/2014/10/Political-Polarization-and-Media-Habits-FINAL-REPORT-7-27-15.pdf. Accessed 21 March 2019.
58. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowd-sourced judgments of news source quality. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2521–2526 (2019).
59. S. Chakrabarti, Nationalism a driving force behind fake news in India, research shows. BBC News, 12 November 2018. https://www.bbc.com/news/world-46146877. Accessed 9 September 2019.