On the Efficacy of Accuracy Prompts Across Partisan Lines: An Adversarial Collaboration

Cameron Martel[1] (iD), Steve Rathje[2] (iD), Cory J. Clark[3,4] (iD),
Gordon Pennycook[5] (iD), Jay J. Van Bavel[2], David G. Rand[1,6,7] (iD),
and Sander van der Linden[8]

[1]Sloan School of Management, Massachusetts Institute of Technology; [2]Department of Psychology,
New York University; [3]The Wharton School, University of Pennsylvania; [4]School of Arts and Sciences,
University of Pennsylvania; [5]Department of Psychology, Cornell University; [6]Institute for Data, Systems, and
Society, Massachusetts Institute of Technology; [7]Department of Brain and Cognitive Sciences, Massachusetts
Institute of Technology; and [8]Department of Psychology, University of Cambridge

Abstract

The spread of misinformation is a pressing societal challenge. Prior work shows that shifting attention to accuracy increases the quality of people's news-sharing decisions. However, researchers disagree on whether accuracy-prompt interventions work for U.S. Republicans/conservatives and whether partisanship moderates the effect. In this preregistered adversarial collaboration, we tested this question using a multiverse meta-analysis ($k = 21$; $N = 27,828$). In all 70 models, accuracy prompts improved sharing discernment among Republicans/conservatives. We observed significant partisan moderation for single-headline "evaluation" treatments (a critical test for one research team) such that the effect was stronger among Democrats than Republicans. However, this moderation was not consistently robust across different operationalizations of ideology/partisanship, exclusion criteria, or treatment type. Overall, we observed significant partisan moderation in 50% of specifications (all of which were considered critical for the other team). We discuss the conditions under which moderation is observed and offer interpretations.

Misinformation is a pressing global issue. Researchers have been exploring the psychological underpinnings of belief in and sharing of misinformation (Pennycook & Rand, 2021; Van Bavel et al., 2021; van der Linden, Roozenbeek, et al., 2021) and testing potential interventions to reduce the spread of misinformation (Gwiaździński et al., 2023; Kozyreva et al., 2022). However, there are competing theoretical perspectives regarding the psychological causes of misinformation belief and sharing and the effectiveness of interventions. These disagreements are difficult to resolve when researchers come from different theoretical traditions or rely on different methodologies (van der Linden, 2022). Furthermore, analyses of the same data sets can yield wildly different conclusions (Breznau et al., 2022).

One way to solve theoretical disagreements such as these is to design adversarial collaborations (Clark & Tetlock, 2023; Mellers et al., 2001), in which researchers with competing hypotheses work together to design a study and preregistered analysis plan to help resolve a scientific debate (Clark & Tetlock, 2023). The current article leverages this approach to help resolve a debate in the field about the efficacy of accuracy prompts across different partisan or ideological groups.

Corresponding Author:
Cameron Martel, Sloan School of Management, Massachusetts Institute of Technology
Email: cmartel@mit.edu

We examine whether shifting attention to accuracy reduces misinformation sharing for people across the political spectrum. Pennycook, Rand, and colleagues (Pennycook, McPhetres, et al., 2020; Pennycook, Epstein, et al., 2021) find that social media contexts focus users' limited attention on factors other than accuracy. Consequently, users share content that they could have identified as inaccurate—and then chosen not to share—had they considered accuracy in advance. A series of lab and field experiments have reported that shifting attention to accuracy (e.g., by asking about the accuracy of an unrelated news headline) improved the quality of people's news-sharing decisions (Pennycook, Epstein, et al., 2021; Pennycook, McPhetres, et al., 2020). Similar findings have been replicated in numerous studies (Arechar et al., 2023; Bhardwaj et al., 2023; Calianos et al., 2022; Capraro & Celadin, 2023; Ceylan et al., 2023; Epstein et al., 2023; Offer-Westort et al., 2022; Organisation for Economic Co-operation and Development, 2022; Rasmussen et al., 2022), although some others have found mixed or nonsignificant results (Gavin et al., 2022; Pretus et al., 2022; Roozenbeek et al., 2021).

The most substantial unresolved debate about accuracy prompts centers on their effectiveness among those on the political right. Pennycook and Rand have argued that accuracy prompts should improve sharing quality to the extent that accuracy discernment is greater than sharing discernment (Pennycook & Rand, 2022b). Thus, although there may be specific items for which accuracy prompts are ineffective for those on the right (or left)—for example, falsehoods that are widely believed by one side (Pretus et al., 2022)—accuracy prompts should be effective for people across the political spectrum in general.

Conversely, Rathje, Van Bavel, and van der Linden have argued that some types of accuracy prompts should be less effective for those on the political right, at least in the United States (Rathje, Roozenbeek, et al., 2022). For example, even when nudged toward accuracy, Republicans may be worse than Democrats at identifying true versus false headlines (e.g., J. Allen et al., 2021; Garrett & Bond, 2021; Imhoff et al., 2022; Lawson & Kakkar, 2022; Rathje, Van Bavel, & van der Linden, 2023; van der Linden, Panagopoulos, et al., 2021), although other studies have found no significant partisan differences in accuracy discernment (e.g., Pennycook, Bear, et al., 2020; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019). This might be due to differences in personality styles between liberals and conservatives (Jost et al., 2018; van der Linden, Panagopoulos, et al., 2021), or different norms, beliefs, and motivations between partisan groups that may lead them to prioritize identity-congruent information over accurate information (Pereira et al., 2018; Rathje,

Statement of Relevance

One proposed intervention for reducing the spread of online misinformation is shifting users' attention to accuracy. Prior research shows that prompting users to consider accuracy improves the quality of people's news-sharing decisions. However, researchers disagree on whether accuracy prompts are effective for those on the political right. This debate has substantial practical implications because Americans on the political right tend to share more misinformation. To help resolve this question, two research teams with conflicting previous research conducted an adversarial collaboration. We found robust evidence that accuracy prompts significantly increased sharing quality for both those on the political right and left. However, we also found that this effect was weaker among Republicans in specifications of particular interest to one of the research teams (although this moderation was not robust across all specifications). Our results have important practical considerations for deploying accuracy prompts in the field.

Roozenbeek, et al., 2022). This debate also has practical implications for how social media platforms and policymakers should best curb misinformation online. Because Americans on the right tend to share more misinformation than those on the left (Garrett & Bond, 2021; Grinberg et al., 2019; Guess et al., 2019; Lawson & Kakkar, 2022; Pennycook & Rand, 2022a; Rathje, He, et al., 2022), if accuracy prompts are only minimally effective for right-leaning participants, accuracy prompts may not be a promising intervention for the audience most likely to share misinformation (see Pretus et al., 2022). Existing empirical evidence is ambiguous regarding accuracy prompts and political orientation. For example, Pennycook, Epstein, et al. (2021) found significant effects among both Democrats and Republicans, and Epstein et al. (2021) found no moderation by conservatism. In contrast, Rathje, Roozenbeek, et al. (2022) conducted a meta-analysis of six experiments using single-headline "evaluation" and "importance" prompts and found that the effect on sharing discernment was moderated by partisanship such that the meta-analytic effect size for Democrats was significantly larger than for Republicans. Further, the effect among Republicans was only marginally significant, leading the authors to conclude that "partisanship matters considerably for the success of this intervention," because accuracy prompts have "little to no effect for U.S. conservatives or Republicans" (p. 1). Finally, Pennycook and Rand

(2022a) meta-analyzed 20 studies that used numerous accuracy-prompt approaches and, separately analyzing convenience versus quota-matched samples, found that the accuracy prompts significantly increased sharing discernment among U.S. conservatives and Republicans across specifications, and the effect was not significantly moderated by conservatism and was moderated only by partisanship in some specifications.

To help reconcile these disparate conclusions, we conducted an adversarial collaboration to examine whether partisanship or ideology moderate the effectiveness of accuracy prompts when pooling all studies considered in Rathje, Roozenbeek, et al. (2022) and Pennycook and Rand (2022a; *N* = 27,828), including a multiverse analysis examining the robustness of any potential moderation effects across accuracy-prompt approaches and operationalizations of political orientation.

Open Practices Statement

All hypotheses and analyses were agreed on by all authors on March 15, 2022 and preregistered at https://aspredicted.org/blind.php?x=JCZ_RNR. Two deviations from our preregistration are specified in the Method section and were agreed on by all authors. Our full data and analytic code are available at https://osf.io/jukx9.

Method

We present the results of an adversarial collaboration in which both sets of authors agreed on a preregistered analysis plan to evaluate the results of 20 data sets collected by Pennycook, Rand, and colleagues and one data set collected by van der Linden and colleagues, all of which used similar designs and operationalizations of ideology and partisanship (*N* = 27,828; 511 different headlines; 601,616 total sharing decisions; for study descriptions, see Table 1). Each study had ethical approval from the Yale University Institutional Review Board, the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects, the University of Regina Research Ethics Board, the Cambridge Psychology Research Ethics Committee, and/or the U.S. Army Human Research Protection Office, and participants provided informed consent.

The two research teams, Martel, Pennycook, and Rand (henceforth known as MPR), and Rathje, Van Bavel, and van der Linden (henceforth known as RVV), agreed on the selection of which studies to include and preregistered analyses to perform. Included studies were not designed, conducted, or initially analyzed by both research teams—the current adversarial collaboration instead involved the preregistration of just study inclusion and analysis specifications. Study selection, analysis, and manuscript drafting were overseen by a third-party moderator (C. J. Clark) from the Adversarial Collaboration Project. RVV had previously analyzed six of these data sets (Rathje, Roozenbeek, et al., 2022) and had not been given access to the participant-level data for the other 15 data sets before this collaboration. MPR had access to and collected these data before the adversarial collaboration, yet none of the reported analyses were undertaken before preregistration.

To help assess the robustness of the results, we conducted a multiverse analysis, which reports all reasonable statistical models rather than only one of many (Steegen et al., 2016). Overall, we present 70 total statistical models, with different exclusion criteria, different measures of political orientation, and different subsets of accuracy-prompt approaches.

MPR's key preregistered hypothesis was that there would be a significant positive effect of treatment on sharing discernment for Republicans/conservatives. RVV's key preregistered hypothesis was that the effect of the accuracy nudge on sharing discernment would be moderated by partisan affiliation (specifically Republicans vs. Democrats—excluding Independents) when looking at the entire sample and that this moderation effect would be relatively robust to different operationalizations of ideology/partisanship and different exclusion criteria. RVV specified that their predictions applied only to two types of accuracy prompts (the evaluation and importance interventions, which prompted users to consider accuracy without providing any additional information) and not to the other accuracy-prompt treatments included in the data (e.g., "media-literacy tips," "social norms," and "evaluation with feedback"; for interventions, see Table 2) because these treatments provided novel information in addition to shifting attention to accuracy.

Experimental designs

In each study, only participants who indicated that they use social media were allowed to participate (there were no inclusion or exclusion criteria based on the types of content people reported sharing online). Participants were presented with a set of actual true and false news headlines taken from social media one at a time and in a random order (these were presented in the format of a Facebook post; for an explanation of the methodology behind headline selection, see Pennycook, Binnendyk, et al., 2021). All false headlines were from fact-checking sites (e.g., snopes.com and factcheck.org), and all true headlines came from mainstream news sources. In most cases, headlines were selected for inclusion on the basis of pretesting; for all experiments using political headlines, the headlines

Table 1. Description of the 21 Experiments Included in the Meta-Analysis

Study	Date	Sample (*N*)	Platform	Topic	Accuracy prompts used	Published?
A	September 2017	847	MTurk	Politics	Evaluation	No
B	October 2017	1,158	MTurk	Politics	Evaluation	Pennycook, Epstein, et al. (2021)
C	November 2017	1,248	MTurk	Politics	Evaluation	Pennycook, Epstein, et al. (2021)
D	March 2019	1,007	MTurk	Politics	Importance; norms; reason; importance + norms + reason	No
E	March 2019	1,210	MTurk	Politics	Evaluation (10×); importance + norms + reason; evaluation (10×) with feedback	No
F	April 2019	1,184	Lucid	Politics	Evaluation (10×) with feedback; importance + norms + reason; evaluation + norms; importance + norms	No
G	May 2019	1,286	Lucid	Politics	Evaluation; importance	Pennycook, Epstein, et al. (2021)
H	September 2019	2,296	MTurk	Politics	Evaluation	No
I	March 2020	855	Lucid	COVID-19	Evaluation	Pennycook, McPhetres, et al. (2020)
J	April 2020	621	MTurk	Politics and COVID-19	Evaluation	No
K	April 2020	444	Lucid	Politics	Evaluation	No
L	April 2020	1,192	Lucid	COVID-19	Evaluation; evaluation (10×) with feedback	Epstein et al. (2021)
M	May 2020	2,081	Lucid	COVID-19	Evaluation; tips; norms	Epstein et al. (2021)
N	May 2020	2,778	Lucid	COVID-19	Tips; norms; tips + norms	Epstein et al. (2021)
O	May 2020	2,616	Lucid	COVID-19	Importance; importance + norms	Epstein et al. (2021)
P	September 2020	820	Lucid	COVID-19	Evaluation; tips	No
Q	September 2020	2,010	YouGov	COVID-19	Evaluation; evaluation with feedback; importance + norms; PSA video	Guay et al. (2022)
R	September 2020	2,015	YouGov	Politics	Evaluation; evaluation with feedback; importance + norms; PSA video	No
S	November 2020	162	Lucid	COVID-19	Evaluation; tips	No
T	December 2020	415	Lucid	COVID-19	Tips	No
U	October 2020	1,583	Respondi	COVID-19	Evaluation	Roozenbeek et al. (2021)

Note: All unpublished studies were conducted by Pennycook, Rand, and colleagues. MTurk = Mechanical Turk.

Table 2. Descriptions of the Accuracy Prompts Used Across Experiments

Accuracy prompt	Description
Evaluation[a]	Participants are asked to rate the accuracy of a single neutral (nonpolitical, non-COVID-19) headline. In some variants, they are shown 10 headlines instead of one; in other variants, they are given feedback on whether their answer was correct. When subsetting analyses on the evaluation treatment, we only include studies in which a single headline was shown without feedback.
Importance[a]	Participants are asked how important it is to them to only share accurate news or to not share inaccurate news.
Norms	Participants are told that most other survey respondents think it is very important to only share accurate news.
PSA Video	Participants are shown a 30-s video (in the format of a "public service announcement," although these words are not explicitly mentioned) reminding them to think about accuracy before sharing.
Reason	Participants are asked how important it is to them to only share news that they have thought about in a reasoned, rather than emotional, way.
Tips	Participants are shown a set of minimal digital-literacy tips; for sample tips, see Epstein et al. (2021).

[a]Preregistered by Rathje, Van Bavel, and van der Linden as pertaining to their claims.

were balanced on partisan lean (i.e., there were equally partisan pro-Democratic and pro-Republican headlines), whereas no attempt was made at political balance for the COVID-19 experiments. Furthermore, headlines were intended to be up to date or relevant when the study was run.

As detailed in Table 1, key dimensions of variation across included studies were the subject pool from which the participants were recruited (convenience samples from Amazon Mechanical Turk, or MTurk; samples from Lucid that were quota-matched to the national distribution on age, gender, ethnicity, and region; or samples from YouGov that used sample matching to select representative samples from nonrandomly selected pools of respondents), the topic of the headlines about which the participants were to make sharing decisions (politics, COVID-19), the specific set of headlines shown (and thus the baseline level of sharing discernment between true vs. false headlines), and the particular set of accuracy prompts used (for a description of each accuracy prompt, see Table 2).

Political-orientation measures

Liberal versus conservative ideology was collected in all 21 experiments. In 17 experiments, participants were asked separately about how socially and economically liberal versus conservative they were using 7-point Likert scales; we averaged the two items to generate an overall ideology measure. In the four remaining experiments, participants were instead asked different versions of political-ideology measures. In two experiments (Studies Q and R; see Table 1), participants were asked

a single question about how liberal versus conservative they were using a 5-point Likert scale (for recent work suggesting this assessment is strongly correlated with the average of the social and economic conservatism measure, $r = .94$, $p < .001$, see Lin et al., 2023). In two experiments (Studies S and T; see Table 1), we used 10-point Likert scales to ask participants (a) the extent to which they thought incomes should be made more equal versus there should be greater incentives for individual effort and (b) the extent to which they thought government should take more responsibility to ensure that everyone is provided for versus people should take more responsibility to provide for themselves, and we averaged the two items to generate an ideology measure. In all experiments, the final measure was then rescaled to the interval [0, 1].

A binary measure of preference for the Democratic versus Republican Party (forced choice, no neutral option) was collected in 19 experiments. In five experiments, this question was asked as a binary forced choice. In 16 experiments, it was asked on a 6-point Likert scale (no neutral option) and then binarized for analysis. We used the Likert measures from these 16 experiments as a continuous measure of preference for the Democratic versus Republican Party. In 18 studies, participants completed a categorical measure of their party affiliation, choosing between Democrat, Republican, Independent, and Other. Finally, we considered a binary measure of whether or not participants reported voting for Donald Trump in the 2016 U.S. presidential election that was collected in 18 experiments.

Because of model convergence and run-time issues, we made several minor deviations from our preregistration.

First, we removed from our analyses a platform covariate (MTurk vs. Lucid/YouGov) and could not test MPR's hypothesis that partisan moderation effects would be larger on MTurk than more representative sampling platforms (for analyses including only studies conducted on Lucid/YouGov, $k = 13$, however, see Supplemental Table S11 in the Supplemental Material available online). We made this deviation to achieve convergence of our multilevel models by simplifying our analyses via removal of a four-way interaction tangential to the primary goals of the adversarial collaboration (for linear mixed-effects model convergence remedies, see Brauer & Curtin, 2018). Second, we did not nest headline items by study when computing random intercepts and slopes at the headline-item level. This deviation was made to increase convergence likelihood by minimizing the presence of headlines with a small number of observations—that is, headlines within individual studies may sometimes have few observations, but repeated headline random effects estimated across studies allow for greater observations per item (and it may also be assumed that differences in headline random effects are unlikely to greatly differ across studies for identical headlines). All preregistration deviations were agreed on by both research teams and the third-party moderator.

Results

Our primary analyses all utilized the same multilevel crossed-effects model structure as preregistered. We predicted sharing intention (normalized 0 to 1) by treatment (0 = *control*, 1 = *accuracy prompt*), headline veracity (0 = *false*, 1 = *true*), partisanship/ideology (0 = *maximally conservative/Republican*, 1 = *maximally liberal/Democrat*), and all two- and three-way interactions. We also included the maximal random-effects structure. Specifically, we included random intercepts for participants nested by study and for headlines, random slopes for headline veracity by participant, and random slopes for treatment, partisanship, and their interaction by headline.

This model structure provided two key quantities for testing our two main hypotheses. First, the coefficient for the interaction between treatment and headline veracity indicated the effect of accuracy prompts on sharing discernment (i.e., sharing more true relative to false headlines) among maximally conservative/Republican participants. Second, the coefficient for the three-way interaction between treatment, headline veracity, and partisanship/ideology indicated the extent to which the accuracy-prompt effect on sharing discernment is moderated by partisanship/ideology—that is, whether accuracy prompts are more or less effective for maximally liberal/Democrat participants relative to maximally conservative/Republican participants.

This model structure also allowed for the calculation of additional coefficients of interest, including (a) the treatment effect on false headlines for maximally conservative/Republican participants, as indicated by the baseline treatment coefficient; (b) the treatment effect on true headlines for maximally conservative/Republican participants, as calculated via a general linear-hypothesis (GLH) test of the sum of the treatment coefficient and the interaction between treatment and headline veracity; (c) the treatment effect on discernment for maximally liberal/Democrat participants, as calculated via a GLH test of the sum of the interaction between treatment and headline veracity plus the three-way interaction of treatment, headline veracity, and partisanship/ideology; (d) the treatment effect on false headlines for maximally liberal/Democrat participants, as calculated via a GLH test of the sum of the treatment coefficient and the interaction between treatment and partisanship/ideology; and (e) the treatment effect on true headlines for maximally liberal/Democrat participants, as calculated via a GLH test of the sum of the treatment coefficient, the interaction between treatment and headline veracity, the interaction between treatment and partisanship/ideology, and the three-way interaction between treatment, headline veracity, and partisanship/ideology. These GLH tests essentially calculated direct treatment-effect coefficient outputs as if our model specified that baseline partisanship was maximally liberal/Democrat (for Democrat effects) or that our baseline veracity was true (for effects on true news sharing). Given that partisanship was scaled from 0 (*maximally conservative/Republican*) to 1 (*maximally liberal/Democrat*), results disaggregated by partisanship/ideology should be interpreted as pertaining to participants who are maximally conservative/Republican or liberal/Democrat.

We conducted this model across a variety of specifications. First, we examined two exclusion-criteria specifications: excluding non-social media users and excluding non-social media users *and* participants who indicated that they never share political news online (for studies with political headlines). Second, we explored five different operationalizations of partisanship/ideology: party identification ("DemRep"; Democrat vs. Republican—excluding Independents), relative preference for the Democratic vs. Republican Party ("DemRep_c"; Likert scale), binary forced-choice preference for Democratic versus Republican Party ("DemRepParty"; no independent/neutral option), political conservatism ("Conservative"; average of Likert scales for social and economic political ideology), and voting for Trump versus not voting for Trump in the 2016 presidential election ("Trump2016"). Third, we examined seven different accuracy-prompt treatment definitions: all treatments, excluding treatments that involved social norms, excluding treatments that involved

Table 3. Effect of Accuracy Prompt on Discernment for Maximally Republican/Conservative Participants (Using a Variety of Specifications)

Exclusion of those who do not share political news	Treatment	DemRep	DemRep_c	DemRepParty	Conservative	Trump2016
No	All	0.030***	0.032***	0.029***	0.031***	0.030***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.005)
No	No norms	0.027***	0.029***	0.027***	0.026***	0.025***
		(0.004)	(0.007)	(0.006)	(0.006)	(0.005)
No	No tips	0.030***	0.032***	0.027***	0.030***	0.028***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.005)
No	No norms or tips	0.028***	0.031***	0.027***	0.026***	0.026***
		(0.005)	(0.007)	(0.006)	(0.006)	(0.005)
No	Evaluation	0.022***	0.022**	0.019**	0.017*	0.020***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.006)
No	Evaluation or importance	0.024***	0.025***	0.022***	0.020**	0.022***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.006)
No	Multiple interventions	0.055***	0.056***	0.050***	0.061***	0.054***
		(0.008)	(0.010)	(0.010)	(0.010)	(0.009)
Yes	All	0.031***	0.034***	0.031***	0.033***	0.030***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.006)
Yes	No norms	0.029***	0.032***	0.031***	0.030***	0.025***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.006)
Yes	No tips	0.030***	0.033***	0.030***	0.032***	0.029***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.006)
Yes	No norms or tips	0.030***	0.035***	0.031***	0.030***	0.026***
		(0.005)	(0.007)	(0.007)	(0.007)	(0.006)
Yes	Evaluation	0.023***	0.027**	0.023**	0.021**	0.019**
		(0.006)	(0.008)	(0.007)	(0.008)	(0.007)
Yes	Evaluation or importance	0.024***	0.028***	0.025***	0.023**	0.022***
		(0.006)	(0.008)	(0.007)	(0.008)	(0.007)
Yes	Multiple interventions	0.057***	0.055***	0.049***	0.061***	0.056***
		(0.010)	(0.012)	(0.011)	(0.012)	(0.010)

Note: Coefficients (multiplied by 100) can be interpreted as percentage-point increases in true relative to false news sharing in accuracy-prompt treatment conditions. Standard errors are provided in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

digital-literacy tips, excluding treatments that involved either social norms or digital-literacy tips, including the single-item evaluation treatment only, including the evaluation or accuracy-importance treatments only, and including treatments with multiple interventions. Altogether, these specifications entailed performing 70 analyses. Note that these 70 model specifications were conducted on highly overlapping data, and the results should not be interpreted as analyses from independent data sets.

We found that accuracy prompts had a significant positive effect on sharing discernment for Republicans/conservatives across all 70 model specifications (minimum $b = 0.017$, $p = .011$; maximum $b = 0.061$, $p < .001$; see Table 3 and Fig. 1). To contextualize the size of this effect in practical terms, these results indicate that accuracy prompts increase true sharing, relative to false sharing, by an additional 1.7 to 6.1 percentage points,

depending on the treatment assessed. These increases in discernment were driven primarily by decreased sharing of false news rather than increased sharing of true news. For example, when assessing the models with the least and greatest discernment effects for Republicans, false news sharing decreased between 3.3% and 14.6% relative to the control, respectively (for full results disaggregated by news veracity, see Tables S3 and S5 and Figs. S1a and S1c in the Supplemental Material). These effects were robust across exclusion of those who do not share political news, different partisanship/ideology specifications, and treatment types and provide evidence in favor of MPR's preregistered hypothesis that accuracy prompts are effective for Republicans/conservatives.

We also found that accuracy prompts had a significant positive effect on sharing discernment for Democrats/liberals across all 70 model specifications (minimum
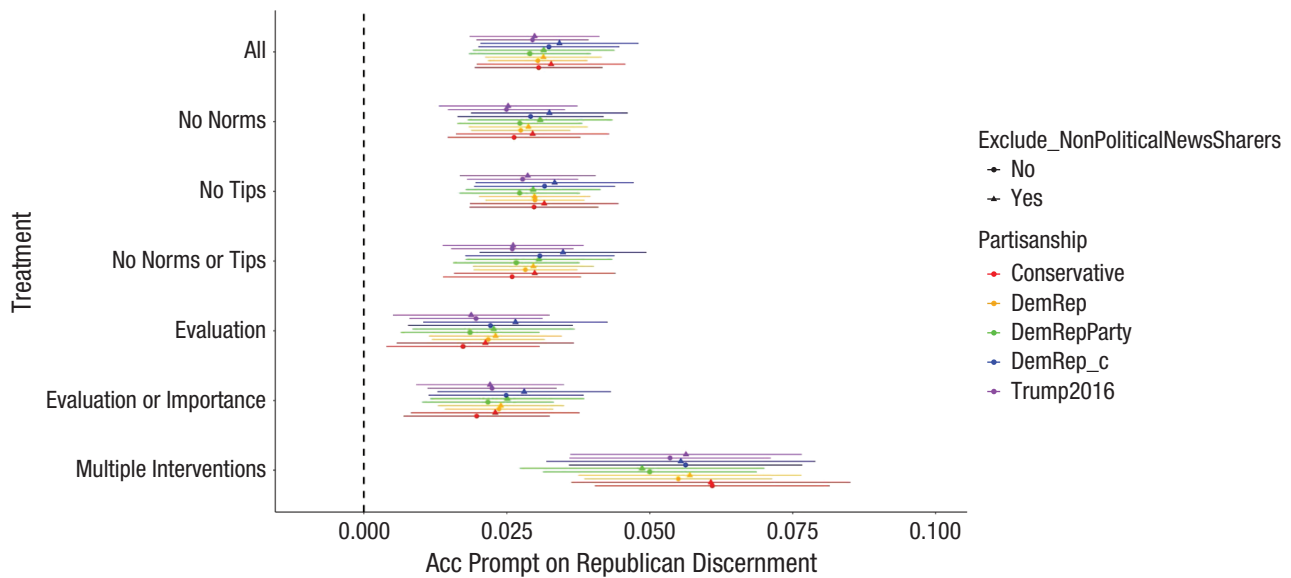
Fig. 1. Coefficient plot of accuracy-prompt effects on discernment for Republican/conservative participants. Coefficients (multiplied by 100) can be interpreted as percentage-point increases in true relative to false news sharing in accuracy-prompt treatment conditions. Coefficients were calculated from the interaction between treatment and headline veracity in the main model. Point estimates reflect discernment coefficients. Error bars reflect 95% confidence intervals.

$b = 0.032$, $p < .001$; maximum $b = 0.065$, $p < .001$; see Table 4 and Fig. 2). Practically, these results suggest that accuracy prompts increase true, relative to false, sharing by an additional 3.2 to 6.5 percentage points, depending on the accuracy-prompt treatment delivered. We again found that these discernment effects were driven by decreased sharing of false news. For instance, in the model specifications with the least and greatest discernment effects for Democrats, false news sharing decreased between 7.6% and 19.1% relative to the control, respectively (for full results disaggregated by news veracity, see Tables S4 and S6 and Figs. S1b and S1d).

We found that 35 models (50%) provided evidence of a significant moderation of accuracy-prompt efficacy by partisanship/ideology across our 70 model specifications (see Table 5 and Fig. 3). In particular, party identification and voting for Trump in 2016 appeared to robustly moderate the accuracy-prompt effect such that accuracy prompts were more effective for Democrats than Republicans or those who reported voting for Trump (12 of 14 model specifications for each partisanship definition; no evidence of moderation when examining "multiple interventions" treatment definition, which refers to cases in which different interventions were stacked together). Binary party preference also moderated accuracy-prompt effectiveness in eight of 14 specifications (six of seven when including those who do not share political news and two of seven when excluding those who do not share political news). However, we found evidence that political conservatism moderated the accuracy-prompt effect in only three of

14 model specifications, and we found no significant moderation by continuous relative preference for the Democratic versus Republican Party in any specification. Interestingly, partisan identity—rather than political ideology—seems to be the more consistent moderator, and binary partisanship specifications were more consistent than continuous ones.

RVV's key specifications of interest were significant. Specifically, RVV's preregistered hypotheses was that effects of the evaluation treatment (i.e., the most common type of accuracy nudge) and the importance treatment would be moderated by partisanship/ideology. When using binary party identification (i.e., excluding Independents) and including participants who reported never sharing political news on social media, which RVV preregistered as their key specification, there was significant moderation when examining studies using the evaluation prompt ($p = .007$) or when examining studies using either the evaluation prompt or the importance prompt ($p = .016$). RVV secondarily preregistered the expectation that this moderation of the evaluation and/or importance treatments would be robust across alternative specifications of political orientation and exclusion criteria. The results were mixed with respect to this prediction: The effect was not significantly moderated by a continuous party-identification measure across any of the relevant specifications and was not significantly moderated by conservatism when excluding participants who reported never sharing political news online; it was, however, consistently moderated by party identification and voting for Trump in 2016 (for histogram of

Table 4. Effect of Accuracy Prompt on Discernment for Maximally Democrat/Liberal Participants (Using a Variety of Specifications)

Exclusion of those who do not share political news	Treatment	DemRep	DemRep_c	DemRepParty	Conservative	Trump2016
No	All	0.043***	0.039***	0.043***	0.044***	0.043***
		(0.004)	(0.005)	(0.005)	(0.006)	(0.003)
No	No norms	0.042***	0.038***	0.044***	0.044***	0.043***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.004)
No	No tips	0.043***	0.039***	0.042***	0.045***	0.041***
		(0.004)	(0.005)	(0.005)	(0.006)	(0.003)
No	No norms or tips	0.043***	0.038***	0.043***	0.047***	0.042***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.004)
No	Evaluation	0.040***	0.033***	0.041***	0.046***	0.038***
		(0.004)	(0.007)	(0.005)	(0.006)	(0.004)
No	Evaluation or importance	0.040***	0.034***	0.041***	0.045***	0.038***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.004)
No	Multiple interventions	0.058***	0.058***	0.055***	0.054***	0.059***
		(0.007)	(0.009)	(0.009)	(0.011)	(0.007)
Yes	All	0.045***	0.041***	0.046***	0.045***	0.045***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.004)
Yes	No norms	0.045***	0.040***	0.047***	0.045***	0.046***
		(0.004)	(0.006)	(0.006)	(0.006)	(0.004)
Yes	No tips	0.044***	0.040***	0.045***	0.046***	0.044***
		(0.004)	(0.006)	(0.005)	(0.006)	(0.004)
Yes	No norms or tips	0.046***	0.040***	0.047***	0.048***	0.046***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.004)
Yes	Evaluation	0.041***	0.032***	0.042***	0.045***	0.041***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.005)
Yes	Evaluation or importance	0.041***	0.034***	0.042***	0.044***	0.040***
		(0.005)	(0.007)	(0.006)	(0.007)	(0.005)
Yes	Multiple interventions	0.062***	0.065***	0.063***	0.06***	0.063***
		(0.008)	(0.011)	(0.010)	(0.012)	(0.008)

Note: Coefficients (multiplied by 100) can be interpreted as percentage-point increases in true relative to false news sharing in accuracy-prompt treatment conditions. Standard errors are provided in parentheses.
*$p < .05$. **$p < .01$. ***$p < .001$.

p values by analysis type, see Fig. 4; for scatterplot of coefficients comparing treatment effect on sharing discernment for Republicans/conservatives vs. Democrats/liberals, see Fig. 5). Interestingly, we found that partisan moderation was more consistently found when assessing political orientation via dichotomous measures rather than continuous measures.

We also conducted several post hoc robustness checks of our analyses. First, we reexamined our effects for the conservatism political-orientation measure when only including studies that used a common political-orientation scale ($k = 17$). In support of MPR's hypothesis, we found that accuracy prompts significantly improved sharing discernment among conservatives under 12 of the 14 specifications (and improved sharing discernment among liberals in all 14 specifications). We also found that conservatism moderated this effect in eight of 14 specifications—and in support of RVV's hypothesis, this was true for all four evaluation and evaluation or importance treatment models (see Table

S10). Second, one potential limitation of the studies included in this multiverse meta-analysis is whether the samples included representative Republican/conservative participants. Therefore, we next reexamined our effects for the evaluation and evaluation or importance models when only including studies that used samples that were more representative (via quota sampling, i.e., studies conducted on Lucid or YouGov only; $k = 13$; see Table 1). In support of MPR's hypothesis, we found in all 10 models that accuracy prompts significantly improved sharing discernment among conservatives/Republicans. We also found this for liberals/Democrats. In contrast to the predictions of RVV, however, we did not find significant moderation by political orientation in any of these models (see Table S11).

Discussion

We provide a comprehensive assessment of whether accuracy prompts improve sharing discernment across
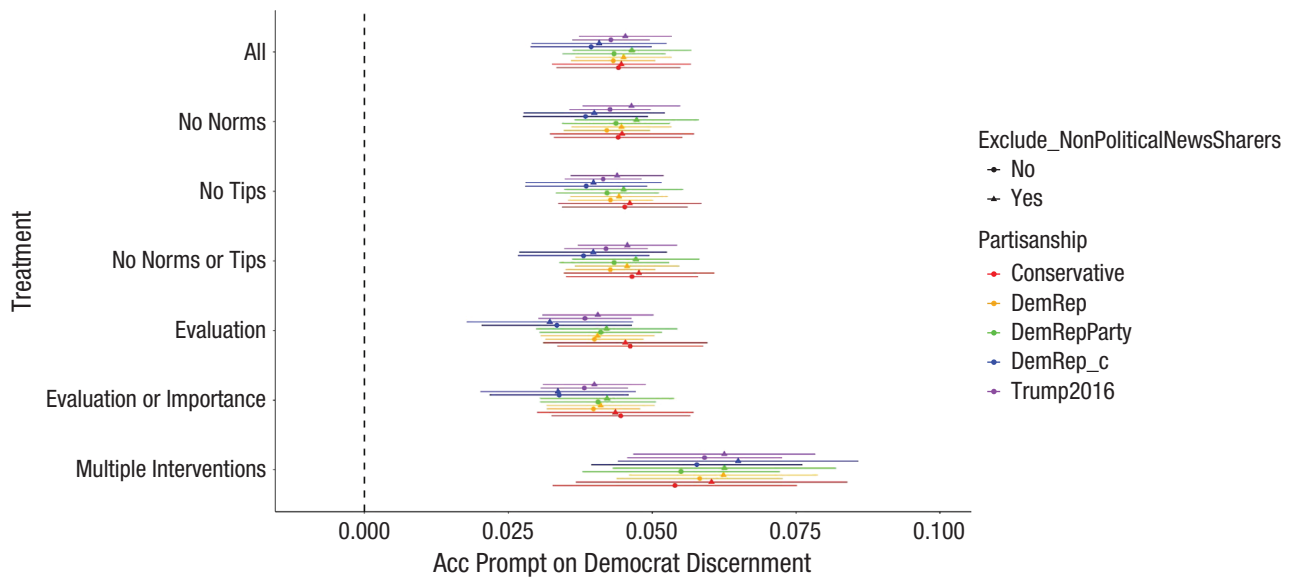
Fig. 2. Coefficient plot of accuracy-prompt effects on discernment for Democrat/liberal participants. Coefficients (multiplied by 100) can be interpreted as percentage-point increases in true relative to false news sharing in accuracy-prompt treatment conditions. Coefficients were calculated from a general linear-hypothesis test of the sum of the interaction between treatment and headline veracity plus the three-way interaction between treatment, headline veracity, and partisanship/ideology. Point estimates reflect discernment coefficients. Error bars reflect 95% confidence intervals.

the political spectrum. The results of our adversarial collaboration indicate two main findings. First, for all 70 model specifications examined, accuracy prompts significantly increased sharing discernment for participants on the political right (by 1.7 to 6.1 percentage points). These results strongly support MPR's key hypothesis that accuracy prompts improve sharing quality among Republicans/conservatives. Second, in half of the 70 models, political orientation significantly moderated accuracy-prompt efficacy such that accuracy prompts were more effective for those on the left than those on the right. The main model specifications preregistered by RVV were significant. RVV's second preregistered hypothesis that their findings would be robust to various measures of partisanship/ideology led to significant results in 13 of 20 cases. These results provide some evidence that accuracy-prompt efficacy is moderated by political orientation—although this moderation was not robust across all treatments, exclusions, or political-orientation measures. Interestingly, moderation effects were more robust when assessing partisanship via party identification or presidential candidate vote (i.e., voting for Trump in 2016) rather than continuous measures of partisanship or conservatism.

Accuracy-prompt effect on Republican/conservative sharing discernment

Both adversarial collaboration teams agree that the results show a robust effect of accuracy prompts on sharing

discernment for participants on the political right: Prompting right-leaning Americans to think about accuracy increased the quality of their sharing intentions. This effect was significant for all 70 model specifications (Fig. 1), supporting MPR's key preregistered hypothesis. However, both teams also agree that this effect was small. Accuracy prompts decreased false relative to true sharing by between 1.7 and 6.1 percentage points. The effect size for Republicans was similar ($d = 0.14$ vs. $d = 0.11$) when replicating the method previously used by RVV to calculate the meta-analytical effect size in their previous meta-analysis. Nonetheless, small effects such as this might be expected given how subtle most versions of the manipulation were (naturally, the longer forms of the manipulation tended to produce larger effects). Moreover, even small effects can have larger impacts when they cascade through social media networks. Future work should investigate the practical significance of the size of the effect and assess the downstream consequences of this magnitude of sharing-quality improvement.

Moderation of accuracy-prompt effect by political orientation

RVV interpretation. Our main preregistered hypothesis was supported. Our specified models, which looked at Democrats versus Republicans as our measure of partisanship and the single-item evaluation as well as the evaluation and importance treatment combined, were significant ($p = .007$ and $p = .016$, respectively).

Table 5. Partisan Moderation of Accuracy-Prompt Effect on Discernment (Using a Variety of Specifications)

Exclusion of those who do not share political news	Treatment	DemRep	DemRep_c	DemRepParty	Conservative	Trump2016
No	All	0.013*	0.007	0.014*	0.014	0.013*
		(0.006)	(0.010)	(0.007)	(0.010)	(0.006)
No	No norms	0.015*	0.009	0.016*	0.018	0.018**
		(0.006)	(0.010)	(0.007)	(0.010)	(0.006)
No	No tips	0.013*	0.007	0.015*	0.015	0.014*
		(0.006)	(0.010)	(0.007)	(0.010)	(0.006)
No	No norms or tips	0.014*	0.007	0.017*	0.021*	0.016*
		(0.006)	(0.010)	(0.007)	(0.010)	(0.006)
No	Evaluation	0.018**	0.011	0.022**	0.029*	0.019**
		(0.007)	(0.012)	(0.008)	(0.011)	(0.007)
No	Evaluation or importance	0.016*	0.009	0.019*	0.025*	0.016*
		(0.006)	(0.011)	(0.008)	(0.011)	(0.007)
No	Multiple interventions	0.003	0.002	0.005	−0.007	0.006
		(0.011)	(0.016)	(0.013)	(0.018)	(0.011)
Yes	All	0.014*	0.007	0.015	0.012	0.015*
		(0.007)	(0.011)	(0.008)	(0.011)	(0.007)
Yes	No norms	0.016*	0.008	0.016*	0.015	0.021**
		(0.007)	(0.011)	(0.008)	(0.011)	(0.007)
Yes	No tips	0.014*	0.006	0.015	0.015	0.015*
		(0.007)	(0.011)	(0.008)	(0.011)	(0.007)
Yes	No norms or tips	0.016*	0.005	0.017	0.018	0.02*
		(0.007)	(0.012)	(0.009)	(0.012)	(0.008)
Yes	Evaluation	0.017*	0.006	0.019*	0.024	0.022*
		(0.008)	(0.013)	(0.010)	(0.013)	(0.008)
Yes	Evaluation or importance	0.017*	0.006	0.017	0.021	0.018*
		(0.007)	(0.012)	(0.009)	(0.012)	(0.008)
Yes	Multiple interventions	0.005	0.010	0.014	−0.000	0.006
		(0.013)	(0.019)	(0.015)	(0.021)	(0.013)

Note: Coefficients (multiplied by 100) can be interpreted as percentage-point increases in sharing discernment for Democrats/liberals relative to Republicans/conservatives in accuracy-prompt treatment conditions. Standard errors are provided in parentheses.
*p < .05. **p < .01. ***p < .001.

We note that the moderation was not significant in every specification (35 of 70 models). We did not consider specifications looking at the social-norms, media-literacy, or long evaluation treatments to be accuracy prompts—and thus did not include these in our preregistered analysis. However, some specifications that were still of interest to us were nonsignificant (such as different measures of ideology/partisanship). Although the moderation effect was less robust than in our prior analysis, we do not find the significant effects entirely surprising given that the effect of accuracy prompts was small and that moderation effects often require roughly 16× the sample size to be detected compared with main effects (Gelman, 2018). Furthermore, we found that the effect size for Democrats ($d = 0.21$) was much smaller than in our earlier analysis ($d = 0.32$), whereas the effect size for Republicans was nearly identical to prior work (see Table S1). Thus, although this effect shows us that the moderation effect was smaller than it was in our earlier analysis, it is primarily smaller not because the effect was stronger for Republicans but because it was weaker for Democrats.

We offer a possible explanation for this moderation effect. The impact of single-item evaluation and importance treatments is known to be moderated by perceived headline accuracy (Arechar et al., 2023; Pennycook, Epstein, et al., 2021): The impact of the nudge on someone's willingness to share a (false) headline becomes smaller the more accurate they believe the misinformation to be and can therefore be expected to work less well (or not at all) both for more persuasive/plausible misinformation and among groups who are more prone to believing misinformation. Because several studies have suggested that U.S. conservatives tend to rate misinformation as more accurate than liberals (e.g., Garrett & Bond, 2021; Pennycook & Rand, 2019), this may reduce the treatment impact. Importantly, people rarely encounter and share blatantly "fake" news headlines (J. Allen et al., 2020; Guess et al., 2019). This calls into question the utility of an intervention that predominantly yields a (modest) reduction in self-reported sharing intentions of implausible false headlines for people who do not believe them to begin with.
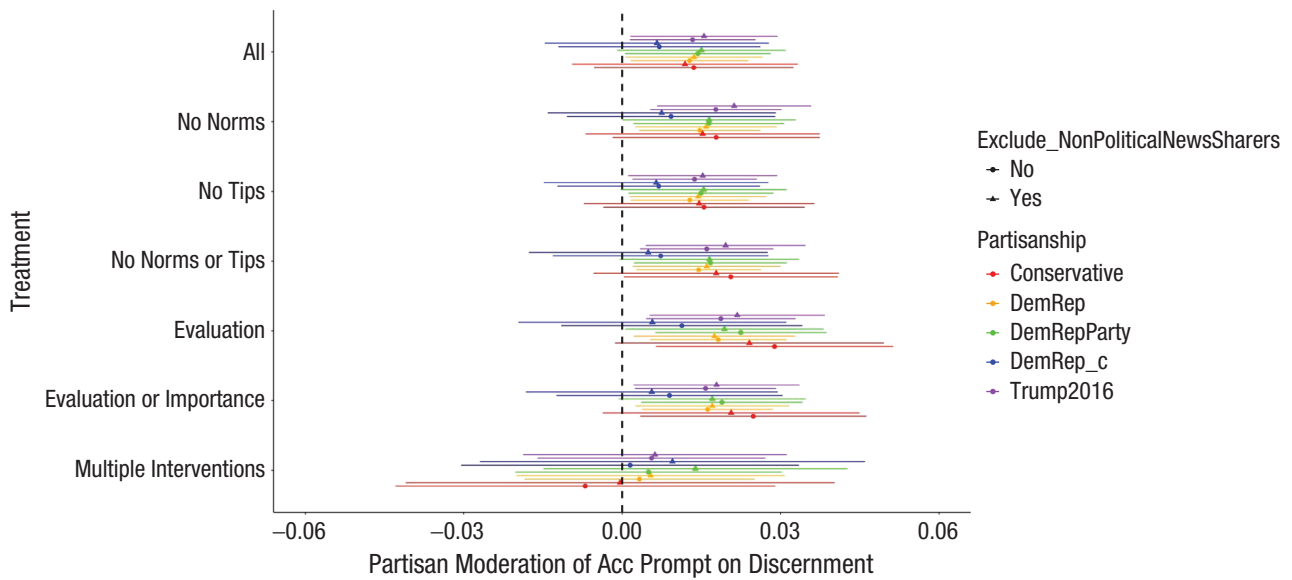
Fig. 3. Coefficient plot of partisan moderation of accuracy-prompt effect on discernment. Point estimates reflect partisanship moderation coefficients. Coefficients (multiplied by 100) can be interpreted as percentage-point increases in sharing discernment for Democrats/liberals relative to Republicans/conservatives in accuracy-prompt treatment conditions. Coefficients were calculated from the three-way interaction between treatment, headline veracity, and partisanship/ideology in the main model. Error bars reflect 95% confidence intervals.

MPR interpretation. We interpret the meta-analytic results as not strongly supporting the claim that partisanship matters considerably for the success of accuracy-prompt interventions. The key interaction between political orientation, treatment, and headline veracity was nonsignificant in half of the 70 specifications and only barely significant in nearly all of the remaining specifications (e.g., only five of the 70 p values were < .01; the statistical
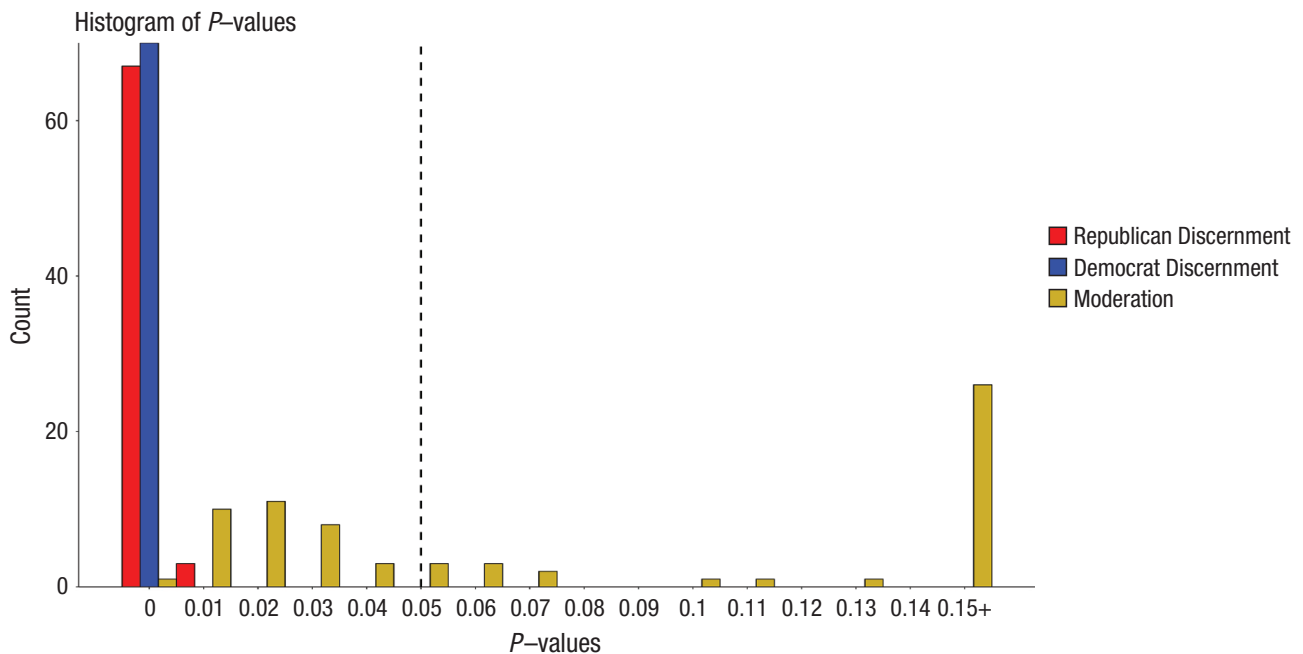
Fig. 4. Histogram of p values by analysis type. The accuracy-prompt effect on discernment for Republicans is shown in red, the accuracy-prompt effect on discernment for Democrats is shown in blue, and partisan moderation of the accuracy-prompt effect on discernment is shown in dark yellow. Histogram bin width = 0.01. Note that this histogram should not be interpreted as a p-curve analysis because it reflects p values from similar analyses from highly overlapping underlying data rather than from independent analyses from separate data sets. The histogram should be interpreted as displaying the robustness of statistical significance across key analyses and specifications in the current multiverse analysis.
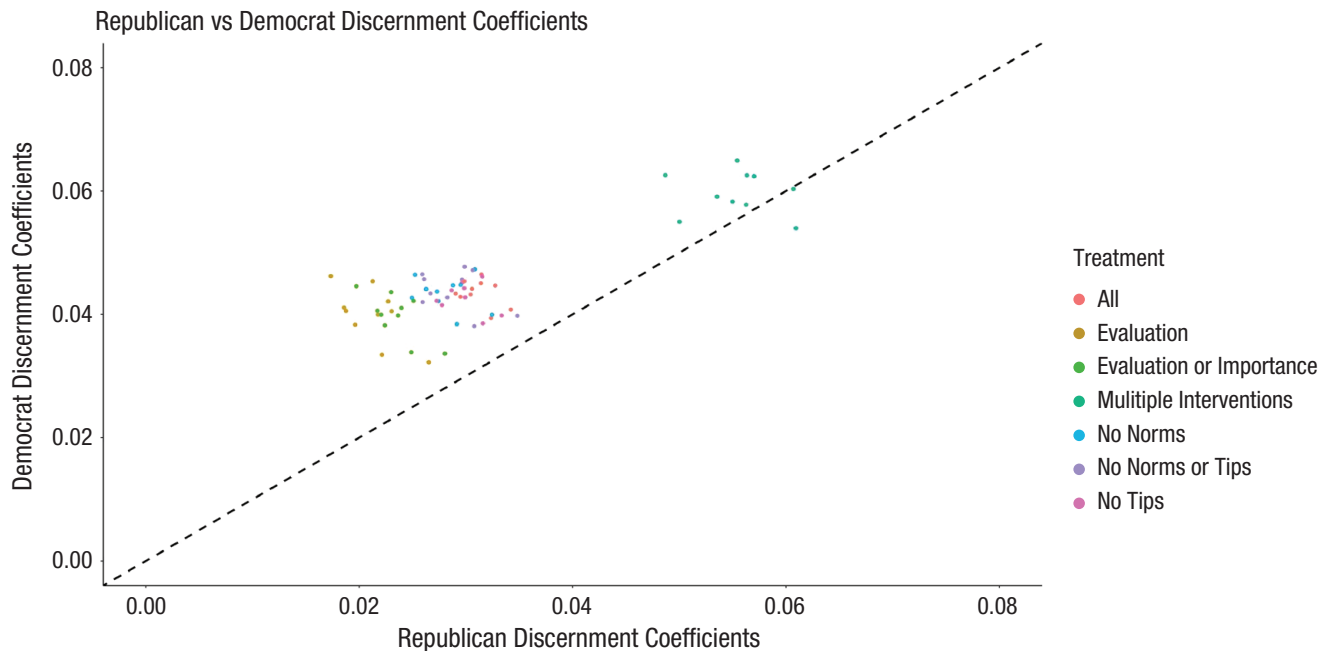
Fig. 5. Scatterplot of Democrat accuracy-discernment coefficients by Republican accuracy-discernment coefficients. The *x*-axis reflects the point estimates of the accuracy-prompt effect on discernment for Republicans/conservatives. The *y*-axis reflects the point estimates of the accuracy-prompt effect on discernment for Democrats/liberals.

significance of partisan moderation was not robust across model specifications; see Fig. 4); these results are similar when only considering RVV's preferred specifications. Given that we found robust evidence that the intervention was successful at improving sharing discernment across the political spectrum in all model specifications, these data do not present compelling evidence that partisanship or ideology mattered considerably for the success of the intervention. That said, our results do suggest that the intervention may work somewhat better for Democrats than Republicans in some cases.

RVV explained the potential moderation by noting that Republicans may be worse at identifying true versus false news—however, existing research has provided mixed evidence (and surely depends on the specific items used; see Clemm von Hohenberg, 2023). Future research could address this directly by investigating whether asymmetries in accuracy discernment explain why accuracy prompts are sometimes more effective for Democrats. Regardless, partisan differences in accuracy-prompt treatment effects vary substantially across headlines. Promisingly, previous meta-analytic estimates have shown that accuracy prompts are significantly more effective for politically concordant headlines (Pennycook & Rand, 2022a)—thus, we expect that accuracy prompts are effective for Republicans when evaluating pro-Republican news content, an important practical consideration given that a larger proportion of misinformation in the United States is shared by conservatives (Guess et al., 2019).

Furthermore, field data from low-quality news sites also show that more blatantly false articles are shared more often (Stewart et al., 2021). Thus, rather than being unimportant, it seems that highly implausible false claims—the claims that are most affected by accuracy prompts—are particularly important to intervene on.

Consensus and future directions. Both adversarial collaboration teams agree that the results provide some evidence of partisan moderation of accuracy-prompt effectiveness such that under some treatments and political-orientation measures accuracy prompts are less effective for Republicans/conservatives compared with Democrats/liberals. This moderation was not strongly robust across all treatments and political-orientation measures, but we almost never observed the reverse effect (stronger effects for Republicans than Democrats). Future work should examine underlying mechanisms for when and why this moderation may exist—for instance, assessing whether moderation may be attributable to partisan differences in accuracy discernment or rather to heterogeneous effects across headlines by political concordance.

Potential limits on generalizability

Our current work has several potential limits on its generalizability. First, all of the studies meta-analyzed here assessed participants recruited from online sampling panels (e.g., MTurk, Lucid, YouGov, Respondi). Although we further assessed our findings when

including only studies that were more representative (quota-sampled respondents; Table S11), our current results address only data generated from participants who in some way opted in to online research studies. The effects of accuracy prompts and their moderation by partisanship could differ for participants who are unlikely to participate in such online studies. Second, the studies assessed here all primarily recruited U.S.-based participants. Future research may examine how accuracy prompts may vary in their effectiveness across the political spectrum in non-U.S. contexts (Arechar et al., 2023). Third, the news stimuli used across all studies were sampled via the procedure outlined in Pennycook, Binnendyk, et al. (2021), in which false stories were selected from those fact-checked by professional fact-checking organizations and true stories were selected from mainstream news outlets. All stimuli were then presented as Facebook-style headlines. Future work may examine the efficacy or partisan moderation of accuracy prompts for headlines that are not so easily designated as true or false (e.g., accurate but misleading or potentially harmful headlines; J. N. L. Allen et al., 2023). Finally, the experiments meta-analyzed here were all online-laboratory experiment-style studies—future work should assess the cross-partisan efficacy of accuracy prompts in field experiments on platforms such as Facebook and X (formerly known as Twitter).

Conclusion

In a novel adversarial meta-analysis of 21 experiments with more than 27,000 participants, we found robust evidence that accuracy prompts significantly increase the quality of content that is subsequently shared by participants across the political spectrum. However, we also found some evidence that this effect may be weaker among Republicans (although perhaps not conservatives). Combating misinformation remains a shared goal among all authors, and adversarial collaborations can help drive such research forward.

Transparency

Action Editor: Yoel Inbar
Editor: Patricia J. Bauer
Author Contributions

David G. Rand and Sander van der Linden are joint senior authors.

Cameron Martel: Data curation; Formal analysis; Investigation; Methodology; Writing – original draft; Writing – review & editing.
Steve Rathje: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Writing – original draft; Writing – review & editing.
Cory J. Clark: Conceptualization; Methodology; Project administration; Resources; Supervision; Validation; Writing – original draft; Writing – review & editing.
Gordon Pennycook: Conceptualization; Data curation; Investigation; Methodology; Resources; Writing – original draft; Writing – review & editing.
Jay J. Van Bavel: Conceptualization; Investigation; Methodology; Project administration; Resources; Supervision; Writing – review & editing.
David G. Rand: Conceptualization; Data curation; Investigation; Methodology; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing.
Sander van der Linden: Conceptualization; Data curation; Investigation; Methodology; Project administration; Resources; Supervision; Writing – original draft; Writing – review & editing.

Declaration of Conflicting Interests

Funding

Open Practices

This article has received the badges for Open Data and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience .org/publications/badges

ORCID iDs

Cameron Martel https://orcid.org/0000-0003-3181-4309
Steve Rathje https://orcid.org/0000-0001-6727-571X
Cory J. Clark https://orcid.org/0000-0002-3083-9179
Gordon Pennycook https://orcid.org/0000-0003-1344-6143
David G. Rand https://orcid.org/0000-0001-8975-2783

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/09567976241232905

References

Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, *7*(36), Article eabf4393. https://doi.org/10.1126/sciadv.abf4393

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), Article eaay3539. https://doi.org/10.1126/sciadv.aay3539

Allen, J. N. L., Watts, D. J., & Rand, D. (2023). *Quantifying the impact of misinformation and vaccine-skeptical content on Facebook*. PsyArXiv. https://doi.org/10.31234/osf.io/nwsqa

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., Gully, A., Lu, J. G., Ross, R. M., Stagnaro, M. N., Zhang, Y., Pennycook, G., & Rand, D. G. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502–1513.

Bhardwaj, V., Martel, C., & Rand, D. G. (2023). Examining accuracy-prompt efficacy in combination with using colored borders to differentiate news and social content online. *Harvard Kennedy School Misinformation Review*, *4*(1). https://doi.org/10.37016/mr-2020-113

Brauer, M., & Curtin, J. J. (2018). Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, *23*(3), 389–411.

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., . . . Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences, USA*, *119*(44), Article e2203150119. https://doi.org/10.1073/pnas.2203150119

Calianos, J., Byles, O., Francis, S., Kot, C. H. B., Seo, H. N., & Nyhan, B. (2022). *The effects of accuracy salience and affective polarization on truth discernment in online news sharing*. Dartmouth College. https://bpb-us-e1.wpmucdn.com/sites.dartmouth.edu/dist/5/2293/files/2021/09/online-news-sharing.pdf

Capraro, V., & Celadin, T. (2023). "I think this news is accurate:" Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news. *Personality and Social Psychology Bulletin*, *49*(12), 1635–1645. https://doi.org/10.1177/01461672221117691

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences, USA*, *120*(4), Article e2216614120. https://doi.org/10.1073/pnas.2216614120

Clark, C. J., & Tetlock, P. E. (2023). Adversarial collaboration: The next science reform. In C. L. Frisby, R. E. Redding, W. T. O'Donohue, & S. O. Lilienfeld (Eds.), *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions* (pp. 905–927). Springer.

Clemm von Hohenberg, B. (2023). Truth and bias, left and right: Testing ideological asymmetries with a realistic news supply. *Public Opinion Quarterly*, *87*(2), 267–292. https://doi.org/10.1093/poq/nfad013

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*, *2*(3). https://doi.org/10.37016/mr-2020-71

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances*, *9*(9), Article eabo6169. https://doi.org/10.1126/sciadv.abo6169

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), Article eabf1234. https://doi.org/10.1126/sciadv.abf1234

Gavin, L., McChesney, J., Tong, A., Sherlock, J., Foster, L., & Tomsa, S. (2022). Fighting the spread of COVID-19 misinformation in Kyrgyzstan, India, and the United States: How replicable are accuracy nudge interventions? *Technology, Mind, and Behavior*, *3*(3). https://doi.org/10.1037/tmb0000086

Gelman, A. (2018, November 15). You need 16 times the sample size to estimate an interaction than to estimate a main effect. *Statistical Modeling, Causal Inference, and Social Science*. https://statmodeling.stat.columbia.edu/2018/03/15/need16

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374–378.

Guay, B., Pennycook, G., & Rand, D. (2022). *Examining partisan asymmetries in fake news sharing and the efficacy of accuracy prompt interventions*. PsyArXiv. https://psyarxiv.com/y762k

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), Article eaau4586. https://doi.org/10.1126/sciadv.aau4586

Gwiaździński, P., Gundersen, A. B., Piksa, M., Krysińska, I., Kunst, J. R., Noworyta, K., Olejniuk, A., Morzy, M., Rygula, R., & Wójtowicz, T. (2023). Psychological interventions countering misinformation in social media: A scoping review. *Frontiers in Psychiatry*, *13*, Article 2872. https://doi.org/10.3389/fpsyt.2022.974782

Imhoff, R., Zimmer, F., Klein, O., António, J. H., Babinska, M., Bangerter, A., Bilewicz, M., Blanuša, N., Bovan, K., & Bužarovska, R. (2022). Conspiracy mentality and political orientation across 26 countries. *Nature Human Behaviour*, *6*, 392–403.

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83. https://doi.org/10.1016/j.copsyc.2018.01.003

Kozyreva, A., Lorenz-Spreen, P., Herzog, S., Ecker, U., Lewandowsky, S., Hertwig, R., Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Maertens, R., Panizza, F., Pennycook, G., Rand, D. G., Rathje, S., Reifler, J., . . . Wineberg, S. (2022). Toolbox of

interventions against online misinformation and manipulation. PsyArXiv. https://psyarxiv.com/x8ejt

Lawson, M. A., & Kakkar, H. (2022). Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General*, *151*(5), 1154–1177. https://doi.org/10.1037/xge0001120

Lin, H., Rand, D. G., & Pennycook, G. (2023). Conscientiousness does not moderate the association between political ideology and susceptibility to fake news sharing. *Journal of Experimental Psychology: General*, *152*(11), 3277–3284. https://doi.org/10.1037/xge0001467

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275.

Offer-Westort, M., Rosenzweig, L. R., & Athey, S. (2022). *Battling the coronavirus infodemic among social media users in Africa*. arXiv. https://doi.org/10.48550/arXiv.2212.13638

Organisation for Economic Co-operation and Development. (2022). *Misinformation and disinformation: An international effort using behavioural science to tackle the spread of misinformation* (OECD Public Governance Policy Papers No. 21). OECD. https://doi.org/10.1787/b7709d4f-en

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, *66*(11), 4944–4957.

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), Article 25293. https://doi.org/10.1525/collabra.25293

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, *592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pennycook, G., & Rand, D. G. (2022a). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*, Article 2333. https://doi.org/10.1038/s41467-022-30073-5

Pennycook, G., & Rand, D. G. (2022b). Nudging social media toward accuracy. *The Annals of the American Academy of Political and Social Science*, *700*(1), 152–164.

Pereira, A., Harris, E., & Van Bavel, J. (2018). *Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news*. PsyArXiv. https://psyarxiv.com/7vc5d

Pretus, C., Javeed, A., Hughes, D. R., Hackenburg, K., Tsakiris, M., Vilarroya, O., & Van Bavel, J. J. (2022). *The Misleading count: An identity-based intervention to mitigate the spread of partisan misinformation*. PsyArXiv. https://psyarxiv.com/7j26y

Rasmussen, J., Lindekilde, L., & Petersen, M. B. (2022). Public health communication decreases false headline sharing by boosting self-efficacy. PsyArXiv. https://psyarxiv.com/8wdfp

Rathje, S., He, J. K., Roozenbeek, J., Van Bavel, J. J., & van der Linden, S. (2022). Social media behavior is associated with vaccine hesitancy. *PNAS Nexus*, *1*(4), Article pgac207. https://doi.org/10.1093/pnasnexus/pgac207

Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J. J., & van der Linden, S. (2022). Letter to the Editors of *Psychological Science*: Meta-analysis reveals that accuracy nudges have little to no effect for U.S. conservatives: Regarding Pennycook et al. (2020). *Psychological Science*. https://doi.org/10.25384/Sage.12594110.v2

Rathje, S., Van Bavel, J. J., & van der Linden, S. (2023). Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour*, *7*, 892–903. https://doi.org/10.1038/s41562-023-01540-w

Roozenbeek, J., Freeman, A. L., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Stewart, A. J., Arechar, A. A., Rand, D. G., & Plotkin, J. B. (2021). *The game theory of fake news*. arXiv. https://doi.org/10.48550/arXiv.2108.13687

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113.

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467.

van der Linden, S., Panagopoulos, C., Azevedo, F., & Jost, J. T. (2021). The paranoid style in American politics revisited: An ideological asymmetry in conspiratorial thinking. *Political Psychology*, *42*(1), 23–51.

van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Traberg, C. S. (2021). How can psychological science help counter the spread of fake news? *The Spanish Journal of Psychology*, *24*, Article e25. https://doi.org/10.1017/SJP.2021.23