Behavioural Public Policy (2026), **10**, 558–574
doi:10.1017/bpp.2024.42

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Shoves, nudges and combating misinformation: evidence on a new approach

Ethan Porter[1], Thomas J. Wood[2], David A. Broniatowski[3] and Pedram Hosseini[4]

[1]George Washington University, USA, [2]Ohio State University, USA, [3]George Washington University, USA and [4]George Washington University, USA
Corresponding author: Ethan Porter; Email: evporter@gwu.edu

(Received 11 December 2023; revised 20 August 2024; accepted 23 August 2024)

Abstract
To what extent can the harms of misinformation be mitigated by relying on nudges? Prior research has demonstrated that non-intrusive 'accuracy nudges' can reduce the sharing of misinformation. We investigate an alternative approach. Rather than subtly reminding people about accuracy, our intervention, indebted to research on the bystander effect, explicitly appeals to individuals' capacity to help solve the misinformation challenge. Our results are mixed. On the one hand, our intervention reduces the willingness to share and believe in misinformation fact-checked as false. On the other hand, it also reduces participants' willingness to share information that has been fact-checked as true and innocuous, as well as non-fact-checked information. Experiment 1 offers proof of concept; Experiment 2 tests our intervention with a more realistic mix of true and false social media posts; Experiment 3 tests our interventions alongside an accuracy nudge. The effectiveness of our intervention at reducing willingness to share misinformation remains consistent across experiments; meta-analysis reveals that our treatment reduced willingness to share false content across experiments by 20% of a scale point on a six-point scale. We do not observe the accuracy nudge reducing willingness to share false content. Taken together, these results highlight the advantages and disadvantages of accuracy nudges and our more confrontational approach.

Keywords: misinformation; nudges; bystander effect; accuracy nudge

Introduction

Misinformation has prompted a variety of potential solutions from scholars, policymakers and activists. Ranging from news literacy campaigns (Guess *et al.*, 2020b; Vraga *et al.*, 2021) to factual corrections (Porter and Wood, 2021; Irving *et al.*, 2022) to crowd-sourcing (Pennycook and Rand, 2019) to 'pre-bunking' (Roozenbeek *et al.*, 2022), these interventions are studied for the capacity to both reduce participants' belief in misinformation and their willingness to share it. One

© The Author(s), 2024. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

class of intervention, modeled on the 'nudge' approach pioneered by Thaler and Sunstein (2008), attempts to curb misinformation through small, non-intrusive changes to the choice architecture of social media platforms (Jahanbakhsh *et al.*, 2021) and/or messages delivered by those platforms (Pennycook *et al.*, 2021). In this paper, we evaluate an intervention that takes a divergent approach. Instead of offering a nudge, our intervention is meant to offer a 'shove': It overtly appeals to social media users to do their part to help resolve the misinformation problem. Just as bystanders who confront social challenges can be turned into problem-solvers (e.g., Cialdini, 1984; Fischer *et al.*, 2011; van Bommel *et al.*, 2012; Abbate *et al.*, 2013), our intervention proposes that direct appeals can turn ordinary social media users into active combatants against misinformation.

Across multiple experiments, we observe our intervention reducing belief in misinformation, as well as the willingness to share misinformation that was fact-checked as false by a non-partisan fact-checking organization. These effects are observed across party lines, across a broad range of misinformation topics and when participants are presented with misinformation in more realistic social media environments. In a competitive trial, we do not observe the more subtle accuracy nudge generating similar effects. However, we also find that our intervention can weaken discernment – that is, just as participants become less willing to share misinformation, they also become less willing to share information fact-checked as true and information so innocuous as to not be fact-checked at all.

Two potential explanations are on offer for our mixed results. The first, consistent with dual-processing, holds that the limited time individuals had to consider the stimuli precluded them from deliberating sufficiently. Bago *et al.* (2020) show that deliberation can reduce the sharing of misinformation. Our treatment offers no opportunity for deliberation, and in doing so, it reduces the sharing of both misinformation and accurate information. The second explanation can be found in the fuzzy-trace theory (Broniatowski and Reyna, 2018), which posits that people prefer to make decisions, such as whether or not to share social media, based on mental representations that encode simple categorical contrasts between decision options that encode motivationally relevant social values. In the case of our intervention, the choice to share boils down to contrast between sharing misinformation, possibly risking social opprobrium, or not sharing misinformation and avoiding this negative outcome. In the absence of unambiguous information regarding whether a *particular* article is true or false, this framing emphasizes the uncertainty associated with sharing and therefore encourages conservative behavior – i.e., not sharing.

By studying an intervention designed to be more aggressive than a subtle accuracy nudge, we are studying an intervention commensurate with the ostensible scope of the misinformation problem. While the prevalence of misinformation can be exaggerated (Guess *et al.*, 2019), it can cause harm, including to public health (Greene and Murphy, 2021; Larson and Broniatowski, 2021) and democratic functioning (Nyhan, 2020). One of the architects of nudges has indicated that nudges are likely inappropriate for certain social challenges, particularly when more aggressive approaches are likely to yield more beneficial results (Sunstein, 2013). Misinformation may be one such challenge. But as our evidence shows, approaching challenges more aggressively than a typical nudge approach has both advantages and disadvantages.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Nudges, shoves and misinformation

The accuracy nudge consists of a 'minimal' (Pennycook *et al.*, 2020, p. 5) reminder about the importance of accuracy, delivered by asking people to evaluate the accuracy of certain items. In their useful review, Pennycook and Rand (2022b) estimate that such nudges can improve individuals' ability to discern fake from non-fake news by 72%. In the race to offer tools to combat misinformation, accuracy nudges have garnered considerable attention among scholars and the media (e.g., Benkelman and Mantas, 2020), while social media companies have implemented 'frictions' meant to nudge users against misinformation (as described in Altay (2022)).

Nudges are not always as effective as their designers intended or as results from controlled experiments suggest. Perhaps owing to a poor understanding of the relevant choice architecture (Sunstein 2017), some fail outright (e.g., Oreopoulos and Petronijevic, 2019). Meta-analyses purporting to show reliable effects (Mertens *et al.*, 2022) have been greeted with statistical skepticism (Gelman, 2022). In a comprehensive comparison between 'nudge units' maintained by governments to implement nudges and academic research that evaluates nudges, DellaVigna and Linos (2022) find that effect sizes observed in studies conducted by the former are typically dwarfed by those found in studies conducted by the latter. The larger effects promised by academic research have not materialized when put into practice by governments, with the difference potentially attributable to publication bias (DellaVigna and Linos, 2022; Maier *et al.*, 2022). To be sure, nudges surely sometimes generate improved outcomes, such as increasing vaccination rates (Milkman *et al.*, 2022) and participation in retirement plans (Madrian and Shea, 2001). However, the brief history of nudging is littered with examples of interventions that appear to work in controlled laboratory settings but do not scale effectively when implemented.[1]

For their part, experiments of accuracy nudges have shown impressive effect sizes. In one study, accuracy nudges doubled participants' ability to separate fact from fiction (Pennycook *et al.*, 2020). The effects have also replicated across multiple studies and contexts (Pennycook and Rand, 2022a). However, some replication efforts have failed (Gavin *et al.*, 2022) or detected smaller effects than earlier papers, with fluctuations based on participants' political orientation (Roozenbeek *et al.*, 2021; Pretus *et al.*, 2023). In a recent adversarial collaboration (Martel *et al.*, 2024), the investigators report that, while accuracy nudges work across partisan lines, they do so somewhat differentially, with some evidence showing that Republicans are less affected by accuracy nudges.

Accuracy nudges, of course, are only one potential response to misinformation. Other scholars have studied news literacy campaigns (Guess *et al.*, 2020b), factual corrections produced by journalists and researchers (Porter and Wood, 2024), corrections produced by peers (Bode and Vraga, 2021), inoculation or 'pre-bunking' (Roozenbeek *et al.*, 2022) and briefer warning labels (Brashier *et al.*, 2021). All have been shown to be effective, albeit to varying amounts and under certain

[1]A broader critique of behavioral research, voiced recently by Chater and Loewenstein (2023), argues that focusing on individuals as the locus for change obscures the systemic nature of certain social problems; this critique is beyond the focus of the present paper.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

conditions (see Prike *et al.* (2023) for a discussion of the specific conditions under which anti-misinformation interventions can fail to achieve their objectives).

In earlier work on reducing false beliefs, researchers found that interventions that 'hit [participants] between the eyes' were most likely to succeed (Kuklinski *et al.*, 2000). The intervention we describe and test in this paper is closer in spirit to that dictum. Our intervention confronts individuals explicitly and asks them to help address the misinformation problem. Given the low absolute levels of misinformation sharing (Guess *et al.*, 2019), most people *are* bystanders to the misinformation challenge. Our intervention relies on canonical insights from Cialdini (1984) who argues that directly targeting individuals can arouse bystanders from their slumber and solve social problems.

In summarizing research on the bystander effect, wherein people respond to an emergency when others are present by *not* responding (Darley and Latane, 1968), Cialdini (1984) recommended that individuals be directly targeted for their ability to help. Rather than hoping against hope that an individual will stand out from the crowd and take action without prompting, Cialdini (1984, pp. 138–139) offers the following:

> My advice would be to isolate someone from the crowd: Stare, speak and point directly at that person and no one else … [that person] should understand that he, not someone else, is responsible for providing the aid; and finally, he should understand exactly how to provide it … pick out *one* person and assign the task to that individual.

Subsequent research has shown that the bystander effect is a common, though not universal, phenomenon (Fischer *et al.*, 2011; Philpot *et al.*, 2020). Research has also identified cases when the effect can be reversed. For example, van Bommel *et al.* (2012) find that elevating individuals' perceived self-importance in a crowd spurs them to be more helpful. Abbate *et al.* (2013) demonstrate that pro-social primes can alleviate the bystander effect. Other successful interventions have focused on the potential for future interaction with other bystanders (Gottlieb and Carver, 1980) and the role of shared affiliations between bystanders and victims (Rovira *et al.*, 2021).

Cialdini was writing about effective responses to personal emergencies in crowded urban environments, such as when someone suffers a stroke in the middle of the street. Social problems, of which misinformation is but one example, have similar properties. Most importantly, although they may be generally aware of social problems and personal emergencies, people may not be sure what they can do to address them. They may count on others to step in and take action; they may be uncertain about the scope of the problem and unclear about what they can do. While misinformation has received widespread media attention, individuals are likely unsure about any potential role they could play in solving the problem.

Our treatment 'shove' was designed to communicate exactly what such a role would look like. In all experiments, it appeared to participants in treatment as follows:

> Misinformation is a serious problem. Around the world, people go on social media to spread fake news and tell lies. Unfortunately, some people believe

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

those lies.
You have the power to make a difference. When you see fake news, you should call it out for what it is: FAKE.

Several features of our intervention are worth remarking upon. Note that the message appeals directly to recipients, communicating that they personally can 'make a difference'. This builds on prior work on the importance of elevating the salience of individual responsibility to challenge the bystander effect (van Bommel *et al.*, 2012). The message also echoes findings on the persuasive power of personal appeals that make explicit reference to the recipient (Rogers *et al.*, 2018) while cuing motivationally relevant social values in order to change behaviors (Broniatowski and Reyna, 2018). Finally, the intervention describes others – not the recipient – as 'believ[ing] the lies' and responsible for the problem. In contrast, recipients are told that they can solve the problem if they so choose. It attempts to empower people, as do some inoculation interventions (Roozenbeek *et al.*, 2022).

Broadly, these features are designed to resemble how Cialdini (1984) recommends that people overcome the 'bystander' dynamic that can arise when a social problem emerges: Rather than deferring to others and counting on them to intervene, our intervention urges individuals to take it upon themselves to take action. Thus, recipients are presented with a decision to make with options that are relevant to the social values that are cued.

In particular, our intervention frames the decision to share as a choice between:

a. Share the news article and potentially spread misinformation (potentially causing a bad outcome)
b. Do not share the news article and certainly do not share misinformation (avoiding a bad outcome)

Our approach is distinguished by our focus on the capacity of individuals to overcome the bystander effect and mitigate misinformation, thus avoiding a bad outcome. While Bode and Vraga (2021) and Pennycook and Rand (2019) offer evidence of the effectiveness of crowd-sourcing to confront misinformation, they do not focus on individuals' capacity to change their behavior and thus address a social problem. Just as people confronting stroke victims on the street may be tempted to pass by and defer to others, so too might people aware of the misinformation challenge respond by continuing forth, unclear about what their contribution could be. A nudge may not be enough.

In contrast, our approach emphasizes a clear categorical contrast between sharing content (and thus possibly spreading falsehoods) and not sharing content, avoiding this negative outcome. Thus, our stimuli contain stark categorical contrasts between decision outcomes that cue motivationally relevant values stored in long-term memory.

The rest of the paper proceeds as follows. In Study 1, we test our intervention exclusively on content fact-checked as false. Observing its effectiveness, in Study 2, we test it on a broader set of stimuli: not only content fact-checked as false but also content fact-checked as true and innocuous, non-fact-checked content. Then,

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

in Study 3, we test our treatment as well as an accuracy nudge on the same diverse set of content tested in Study 2. We conclude by discussing the implications and limitations of our present work.

Experiments

Study 1

Study 1 ($n = 2{,}971$) was conducted in December 2021 on Amazon's Mechanical Turk. After answering pretreatment questions (demographics, party ID and the Cognitive Reflection Test), as well as an attention check (modeled on Berinsky *et al.* (2014)), participants were randomized into either treatment or control. Subjects were randomly assigned with equal probability to either treatment or a pure control – that is, they were either exposed to the treatment and then answered outcome questions, or they only answered outcome questions.[2] After randomized exposure to treatment or a pure control, all participants were then presented with six social media posts in random order. Three of the posts had been fact-checked as false by the non-partisan fact-checking organization PolitiFact; the remaining three had been fact-checked as true by the same organization. The three false items were selected because they recently had been fact-checked by the organization (enhancing the ecological realism of the study) and because of their political balance. One item denigrated Democrats; another denigrated Republicans; a third was entirely apolitical. The three tested items, which are fact-checked as false, can be glimpsed in Supplementary Figure A1. Summaries of the items appear in Supplementary Table A1. Similar to Pennycook *et al.* (2020), we restrict our analysis to those who report having a social media account. We also omit respondents who failed our pretreatment attention check.

To measure the effects, after exposure to each post, participants were asked two questions. Following practices in previous work, to assess willingness-to-share, we asked participants 'If you were to see the above article on Facebook, Instagram or other social media, how likely would you be to share it?' Responses ranged from 'extremely likely' to 'extremely unlikely' on a six-point scale. To assess belief accuracy, participants were asked 'To the best of your knowledge, how accurate is this statement?' and then presented with a sentence summarizing the content of the post. For example, under a post alleging that Hunter Biden, Joe Biden's son, had been arrested by the military, participants were asked to assess the accuracy of the following sentence: 'The U.S. military has arrested Hunter Biden'. Potential responses ranged from 'Not at all accurate' to 'Very accurate' on a 1–4 scale.

Supplementary Table A2 offers descriptive statistics and provides evidence of balance across conditions. The average participant was 39.03 years old (SD = 11.7), had nearly obtained a college degree (7.965, with 8 representing completion of a degree; SD = 1.67), and was a weak Democrat (3.10, with 3 representing standing for not strong Democrat; SD = 2.40).

[2]Across studies, we alternated between pure controls and placebos, out of the possibility that placebo selection may inadvertently bias treatment effect estimates, as discussed in Porter and Velez (2021). Our point estimates remain consistent when compared to either placebos or control; however, as discussed below, the placebo sometimes yields surprising effects.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

As described in our pre-registration document, to analyze effects, we created two indices, one for belief accuracy and one for sharing discernment (with the latter understood as willingness to share the false items). (The pre-registration document is available in the Appendix.) The indices were constructed by averaging outcome data pertaining to the three false items. Supplementary Table A3 presents results on belief accuracy. As pre-registered, throughout this paper, we estimate effects via ordinary least squares (OLS) with robust standard errors, sans covariates, with binary variables standing in for condition assignment. On this 1–4 scale, the treatment reduced belief in false items by −0.14 (95% CI: −0.21 to −0.08). Supplementary Table A4 reports results for sharing intention. On a 1–6 scale, the treatment reduced willingness to share false items by −0.25 (95% CI: −0.37 to − −0.13).

Prior research offers reason to believe that these effects may have varied by respondents' political partisanship and their score on the cognitive reflection test (CRT). We measured partisanship via the standard 7-point scale (with responses ranging from 'Strong Democrat' to 'Strong Republican') and CRT via three questions (the total cost of a bat and ball, the time required to make 100 widgets and the growth rate of a lily pad). In Supplementary Tables A5 and A6, we present results that account for a partisanship interaction. Consistent with our pre-analysis plan, we interact treatment assignment with a dummy variable for Republican affiliation. We do not observe Republican affiliation interacting with either the sharing or belief accuracy outcomes. We repeat this exercise for responses to CRT. Following our pre-analysis plan, we trichotomize responses and create dummy indicators, which, in turn, interact with treatment assignments. Supplementary Tables A7 and A8 present results. We find no evidence that CRT interacts with responses to our treatment.

Study 2

Next, we investigate whether our intervention reduces intent to share in more realistic environments. While misinformation attracts deserved attention, it represents only a small fraction of the kinds of content people are exposed to in real-world social media environments (Guess *et al.*, 2020a). With this in mind, the second study exposed participants to a larger and more diverse set of content. Across 16 items, participants were exposed, in random order, to three kinds of items: misinformation fact-checked as false, one item fact-checked as true, and 12 items *not* fact-checked at all. (In Study 1, our pre-analysis document only included analysis plans for the first category.)

This last category was particularly important for realism reasons. Most content on social media is not subjected to review by professional fact-checkers (see Guess *et al.* (2020a) and Guess *et al.* (2019) for evidence of typical content consumed on social media). To develop our set of non-fact-checked content, we relied on CrowdTangle for data on viral items on Facebook in the U.S. at the time the test was fielded. Items in this category included posts featuring a Peanuts cartoon; a picture of Bruce Lee and an inspirational saying. At the time, this experiment was administered, posts like these, not egregious misinformation, represented the most viral items on Facebook. In total, we included 12 non-fact-checked viral items. Supplementary Figure A2 provides examples.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

We also included three items that had been fact-checked as false, with our typical mix of one flattering Democrats, one flattering Republican and one non-partisan. The false item that flattered Democrats claimed that Ivanka Trump had been arrested; the false item that flattered Republicans said that Joe Biden had ceased domestic gas production; the non-partisan item presented an obituary of a Ukrainian war hero. Finally, we included an item fact-checked as true, in which it was conveyed that Southwest Airlines would no longer terminate employees who failed to comply with a COVID-19 vaccine mandate. Supplementary Figures A7–A10 display this stimuli; Supplementary Table A1 summarizes the stimuli used here and in the other studies as well.

The study was conducted on Mechanical Turk in March 2022 ($n = 2{,}481$). After consenting, answering pretreatment questions and passing an attention check, subjects were randomized with equal probability into either a treatment condition, a placebo or a control. The placebo consisted of a nonpolitical news article used in prior research (Nyhan *et al.*, 2019); those in the control group only answered outcome questions. The average respondent was 37.37 years old (SD: 11.01), had obtained a college degree (8.19, with 8 representing completion of a degree; SD: 1.65) and was a weak Democrat (3.17, with 3 representing standing for not strong Democrat; SD: 2.55). More descriptive statistics can be glimpsed in Supplementary Table A9.

For this study, we only measured the effects on willingness to share, as we feared it would have dampened realism and possibly elicited demand effects to repeatedly ask participants to share their beliefs about the innocuous non-fact-checked items. Sharing intention was measured on a 1–6 scale, with choices ranging from 'Extremely unlikely' to 'Extremely likely'.

As Supplementary Table A10 shows, the treatment significantly reduced intent to share false content, by 0.16 points on our scale (95% CI: −0.31 to −0.02). The placebo had no similar effects ($\hat{\beta} = 0.04$; 95% CI: −0.11 to 0.18). In Supplementary Table A11, we report sharing intention among non-fact-checked articles – in other words, among articles that resemble the lion's share of viral content encountered on social media. We do not find our treatment reducing willingness to share such content ($\hat{\beta} = -0.07$; 95% CI: −0.19 to 0.05). However, as Supplementary Table A12 shows, we do observe participants exposed to our treatment being less willing to share a post that had been fact-checked as true by 0.20 on our scale (95% CI: −0.35 to −0.04). Spillover between our treatment's effects on false and non-false items does occur, but only on a select kind of content. The top row of Figure 1 depicts treatment effects for different content types.

As before, we also investigate interactions between our treatment and partisanship, as well as our treatment and responses to CRT. Supplementary Tables A14 and A13 show the results. Here, we find some evidence of partisanship moderating treatment effects, with Republicans adversely impacted by the treatment. While we do not observe this in either of the other studies, this result leads us to believe that partisans may sometimes respond differently to our treatment.

Study 3

In a final study, we evaluate both our treatment as well as an accuracy nudge. In doing so, we rely on the same stimuli from Study 2 (summarized in Supplementary

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Table A1). Participants ($n = 4{,}900$) were recruited in August 2022 over Mechanical Turk. Upon the completion of demographic questions and passage of an attention check, they were assigned with equal probability to either our treatment; an accuracy nudge; the same placebo as Study 2 or a control, wherein participants only answered outcome questions. The tested accuracy nudge instructs participants to rate the accuracy of two social media posts on a 1–5 scale, with one post real and implausible and the other fake and implausible. The posts are taken directly from the accuracy nudge used in Study 3 by Pennycook *et al.* (2021); they can be seen in Supplementary Figures A3 and A4.

The average participant was 37.67 years old (SD: 11.67), had obtained a college degree (8.22, with 8 representing completion of a degree; SD: 1.65) and was a weak Democrat (3.31, with 3 representing standing for not strong Democrat; SD: 2.57). Supplementary Table A15 offers more descriptive statistics and evidence of balance.

Based on evidence from prior studies, we expected our treatment to once again reduce belief in the false items. Whether or not our treatment would be more effective than an accuracy nudge remained an open question.

As before, we analyze the effects on both false and non-false content by taking the average effects for both kinds of items. Table 1 presents results for false content. The effect of our treatment is sharply negative, reducing participants' average willingness to share false viral content by 0.20 points on a six-point scale (95% CI: −0.32 to 0.08). By contrast, the effects of the accuracy prompt on willingness to share false content are not significant, while the point estimate is less than one-third the size of the point estimate generated by our treatment. Even the point estimate from the placebo content is also larger than that from the accuracy nudge. The 0.14 difference between the accuracy nudge and our treatment coefficients is not only sizable but significant ($p < 0.05$, two-tailed).

The story is more complicated when we examine the effects on willingness to share non-false content. As before, we look at the average effect of willingness to share

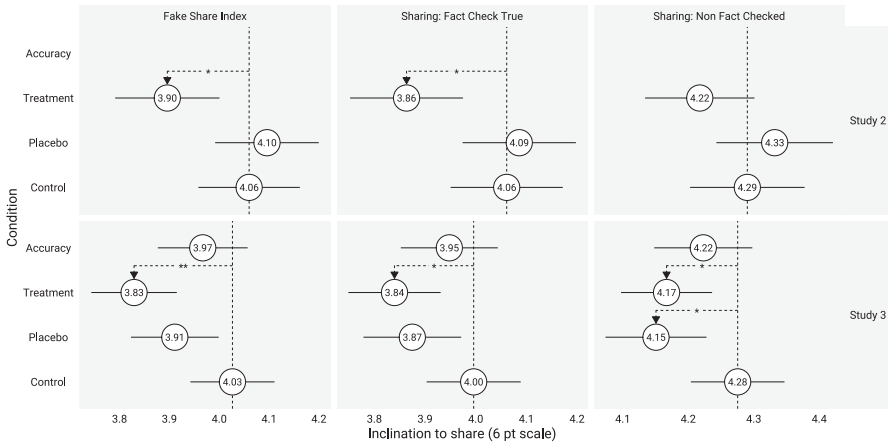


Figure 1. Conditional effects on sharing, studies 2 and 3. *p < 0.05; **p < 0.01.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

across our non-false yet viral items. Our treatment reduces willingness to share *non-false* content, this time by 0.11 (95% CI: −0.21 to −0.01). This is roughly half the size of the reduction observed for false items, suggesting that our treatment leads people to be *somewhat* discerning between false and non-false content. The effects of the accuracy nudge on people's willingness to share non-false content, however, are not significant, and the point estimate is half the size of the point estimate for our treatment. Table 2 displays the results. A similar pattern emerges when we focus on the effects on willingness to share items fact-checked as true. As Supplementary Table A16 shows, the treatment reduced the sharing of this item by 0.16 (95% CI: −0.29 to −0.03) along a 1–6 scale; the accuracy nudge did not ($\hat{\beta} = -0.05$, 95% CI: −0.18 to 0.09). The bottom row of Figure 1 depicts conditional treatment effects for different content types. Again, while only our intervention reduces willingness to share content fact-check as false, we also find it reduces willingness to share content fact-checked as true and content not fact-checked at all.[3]

As usual, we inspect the extent to which any effects were conditional on partisanship and levels of cognitive reflection. We do not observe evidence for either. Results appear in Supplementary Tables A17 and A18.

Discussion

We present evidence documenting the effectiveness of a novel intervention – a shove that draws explicit attention to the social challenge of fake or otherwise misleading news – to combat misinformation. With three pre-registered experiments, we observe our intervention reducing willingness to share and believe in misinformation. While other scholars have shown that gently 'nudging' individuals to be more mindful of accuracy can reduce the willingness to share misinformation, our evidence indicates that aggressively targeting individuals can also do the trick. Our treatment urges social media users to be active participants in combating misinformation. Our results show that they generally comply with positive effects on the factual accuracy of their beliefs and their sharing discernment but with spillover effects on the sharing of non-false content. In no study do we observe CRT scores interacting with our treatment; in two of the three, we do not observe interaction effects by partisanship, either. To understand our results in aggregate, we administer a meta-analysis with random effects for the treatment effects on sharing false content for all three studies. We find that, across studies, our treatment reduced willingness to share by −0.21 on a six-point scale ($p < 0.05$, two-tailed). As Figure 2 shows, we observed consistent effects across studies.

Consider our results from the perspective of a resource-constrained policymaker. Whether working in government or a social media company, assume that this policymaker is interested in reducing the spread of misinformation, as well as believing in it. The policymaker can either use the accuracy nudge which, as commonly studied

[3]The careful reader will also note that the placebo condition also moves outcomes in the same direction as our treatment and the accuracy nudge. While our treatment is the only condition to reduce willingness to share misinformation, both our treatment and the placebo reduce willingness to share content not fact-checked at all. To us, this suggests that responses to sharing queries about these innocuous items may not be stable, in the same way that responses to questions about belief in false claims are not stable (as described in Graham (2023)).

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Table 1. Study 3: effects on sharing fake content

Term	Estimate	Std. error	p-value	conf.low	conf.high
1. Treatment	−0.20	0.06	0.00	−0.32	−0.08
2. Placebo	−0.12	0.06	0.06	−0.24	0.01
3. Accuracy nudge	−0.06	0.06	0.34	−0.18	0.06
4. Intercept	4.03	0.04	0.00	3.94	4.11

Table 2. Study 3: effects on sharing non-fact-checked content

Term	Estimate	Std. error	p-value	conf.low	conf.high
1. Treatment	−0.11	0.05	0.03	−0.21	−0.01
2. Placebo	−0.12	0.05	0.02	−0.23	−0.02
3. Accuracy nudge	−0.05	0.05	0.32	−0.16	0.05
4. Intercept	4.28	0.04	0.00	4.20	4.35

(Pennycook *et al.*, 2021), requires individuals to answer survey questions. Survey completion rates on social media are quite low (Zhang *et al.*, 2020); disseminating the treatment across large numbers of users would thus be cumbersome and costly. In comparison, our intervention can simply be presented to all users, with no survey necessary. When it comes to ease of application, our intervention is thus similar in spirit to the news literacy intervention studied by Guess *et al.* (2020b): All users can be compelled to read it, with no additional action required. And as shown, our intervention is likely to have a larger impact on sharing discernment. If the policymaker wishes to achieve her objectives in their most limited form, our intervention may be less costly than an accuracy nudge. Ultimately, like other anti-misinformation tools (Jigsaw, 2023; Lin *et al.*, 2024), our intervention could be directly presented to social media users while they spend time on the platform. Users could encounter such interventions at random intervals throughout their time on platforms, while sharing, composing content or just browsing. This small friction would likely positively affect a range of outcomes that social media companies purport to care about.

The results of Study 3 should not be understood to suggest that accuracy nudges are always bound to fail. Although we were unable to detect effects on sharing from accuracy nudges in this experiment, the evidence available in the published literature leads us to believe that accuracy nudges can sometimes work. To be clear, however, our failure to find an effect in Study 3 is not owed to low power; we had 99% power to detect very small effects and 100% power to detect small to medium effects. If we regard misinformation as requiring an 'all-hands-on-deck' response, different interventions may be deployed for different ends (consistent with Bak-Coleman *et al.* (2022)). To achieve larger effects on sharing discernment than a typical accuracy nudge, our intervention may be used. But as we show, our intervention comes with costs. Not only is it intentionally more confrontational, but it does somewhat reduce willingness to share *true* content as well as entirely innocuous content. In

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

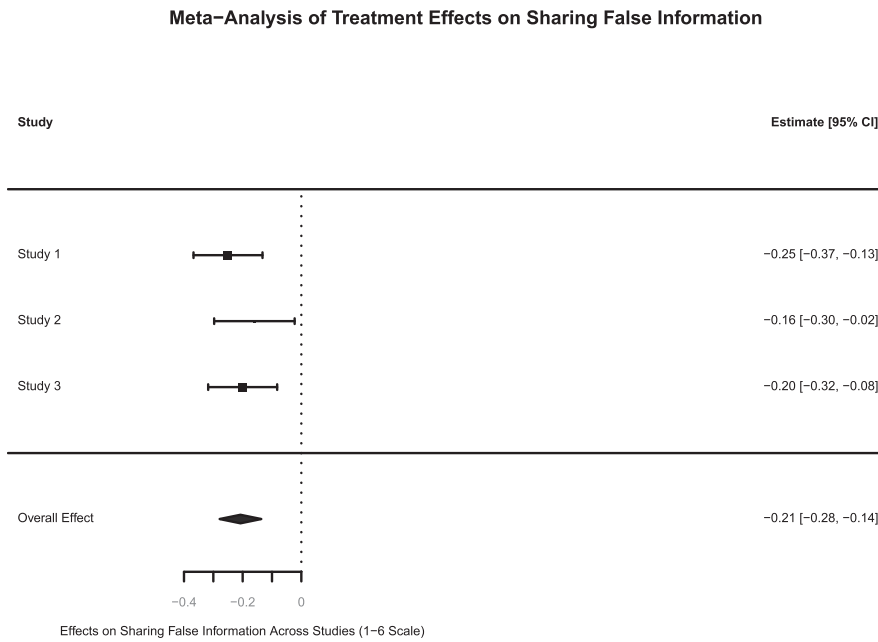


Figure 2. Meta-analysis of our intervention.

contrast, accuracy nudges do not always achieve their intended objective but also have smaller spillovers too. Our intervention is similar to common anti-misinformation tools that reduce belief in falsehoods while simultaneously increasing skepticism toward true information (Hoes *et al.*, 2024).

Standard dual-process theories of reasoning posit that the very factors that are responsible for the advantages of our intervention are also responsible for its disadvantages. According to these theories, our intervention draws people's attention to the problem of misinformation, but its emphatic nature limits the opportunity for deliberation about what constitutes misinformation. As Bago *et al.* (2020) show, offering people the opportunity to deliberate, in ways our intervention does not, reduce belief in inaccurate information *without* reducing belief in accurate information. To be confronted with one's status as a bystander is to be moved to action without invoking deliberative processes – and a lack of deliberation can lead people astray. This explanation echoes Orchinik *et al.* (2023), which finds that people's perceptions of the proportion of true and false claims posted online can shift rapidly.

Another explanation for our mixed findings can be found in the fuzzy-trace theory (Broniatowski and Reyna, 2020), which would view our treatment as offering a stark categorical contrast that consistently favors not sharing (because sharing always contains some risk of spreading misinformation). This, in turn, reduces the sharing of both types of information. Our intervention renders the sharing of misinformation worse than the consequences of not sharing true information, which leads people to take the conservative route and share less of everything. If a version of our intervention were targeted at an issue about which more people have firmer, false beliefs, we would

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

not have much confidence that it would work. For example, a Cialdini-inspired intervention delivered to committed members of the anti-vaccine community would likely not succeed. The strength of pretreatment beliefs and behaviors (Druckman and Leeper, 2012; Howe and Krosnick, 2017) likely conditions the effectiveness of our intervention. This account suggests that interventions explaining *why* a specific stimulus is misinformative may be even more effective when combined with our 'shove' intervention. Such interventions, focused on changing mental representations, maybe more useful (Reyna *et al.*, 2021) and, because they are based on building understanding beyond the scope of a specific stimulus, may be more effective across contexts.

We also wish to offer a note of caution about our stimuli. We used a smaller set of stimuli than some other papers in this field have used (e.g., Pennycook *et al.*, 2021). There may have been variations in the believability and credibility of various headlines, which, in turn, could be impacting our results. Furthermore, all of our stimuli are also dependent on the decisions of fact-checking organizations, which, of course, cannot evaluate every possible viral post. That having been said, PolitiFact does decide which stories to fact-check based on data about popularity (as described in Coppock *et al.* (2023)), so we do believe our tests concern reasonably popular social media posts. For the sake of reaching generalizable conclusions, additional research on a wider set of stimuli is sorely needed. For now, we offer summaries of all tested stimuli in Supplementary Table A1; readers may make their own evaluations of the believability and credibility of our tested items.

There are three additional limitations of the present study worth remarking upon. First, similar to many studies of accuracy nudges, we do not present evidence on the longitudinal effects of our intervention. As with fact-checking, we expect the effects of our treatment to attenuate but remain detectable with sufficiently powered samples (Porter and Wood, 2021). Second, we cannot speak whether the effects would be obtained on a more representative sample. While prior research has shown that misinformation can be profitably studied on Mechanical Turk (Pennycook *et al.*, 2021), with observed effects comparable to those observed in more representative studies (Nyhan *et al.*, 2019), more representative samples would offer important further evidence about our intervention. We also cannot rule out demand effects. While Mummolo and Peterson (2019) offer a good reason to think that demand effects are much less common than assumed, it is possible that participants in our experiments modified their responses in order to satisfy their perceptions of the researchers' wishes.

These limitations set the stage for future research. Future research should systematically test how different interventions can be used in concert and targeted toward specific individuals and groups. It may be the case that some users would benefit more from accuracy nudges than our intervention, and vice versa. Future experiments using our intervention would also help clarify the replicability of the heterogeneous effect results reported in this paper; are the positive effects consistently concentrated among Democrat and independents, as Study 2 would suggest, or are they largely uniform across party lines, as Studies 1 and 3 would indicate? Simply put, answers to these questions require a larger set of studies, perhaps even similar in number to that evaluated by Martel *et al.* (2024). Future research should also conduct panel studies on anti-misinformation interventions to measure whether their effects persist. If the effects of our intervention endure after exposure, we would regard this as evidence against demand effects.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Last but not least, we would be especially enthused about experiments that modify elements of the treatment text; for example, it may be the case that simultaneously encouraging individuals to call out false information and to promote true information would have the same positive effects we observe but with fewer downsides.

For now, however, we have offered consistent evidence in multiple experiments that directly soliciting individuals' help in resolving the misinformation challenge can reduce willingness to believe in and share misinformation. Like bystanders in a crowd asked to do their part to help an ailing individual, our participants do their part to help reduce the spread of, and belief in, misinformation.

Conclusion

With insights gleaned from the bystander literature (e.g., Cialdini, 1984), we show that aggressively confronting individuals about their ability to fight misinformation reduces their willingness to share and believe in misinformation. This effect is consistent across three experiments; in a third experiment, our intervention again reduces the sharing of misinformation, while the tested accuracy nudge does not lead to any such improvements. However, our intervention comes with a downside in that it also reduces the willingness to share accurate information and innocuous information. Our results would likely have been different had we offered participants more time to deliberate about the stimuli (or if we had changed the stimuli). Emphasizing the downside risks of online sharing, as our intervention does, presents individuals with a choice between sharing, and potentially incurring some social risk or not-sharing and therefore incurring no social risk (Broniatowski and Reyna, 2020). In our studies, this choice leads subjects to behave more conservatively and share less overall.

Supplementary material. To view supplementary material for this article, please visit https://doi.org/10.1017/bpp.2024.42.

Data availability statement. Data and code for Studies 1–3 can be found at OSF here: https://bit.ly/3reBO38.

Acknowledgements. We thank Robert Brauneis and Dawn Nunziato. All mistakes are our own.

Authors' contributions. E.P., T.J.W., D.A.B. and P.H. conceived of the research, analyzed data and wrote the paper.

Funding statement. This research was supported by the Social Science Research Council and the New America Foundation and approved by the George Washington University IRB (#NCR202850).

Competing interest. The authors declare no competing interests.

Ethics approval and consent to participate. This research was approved by the George Washington University IRB (#NCR202850).

Consent for publication. All participants provided informed consent before submitting.

References

Abbate, C., S. Ruggieri and S. Boca (2013), 'The effect of prosocial priming in the presence of bystanders', *Journal of Social Psychology*, **153**(5): 619–622.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Altay, S. (2022), How effective are interventions against misinformation? https://doi.org/10.31234/osf.io/sm3vk

Bago, B., D. G. Rand and G. Pennycook (2020), 'Fake news, fast and slow: deliberation reduces belief in false (but not true) news headlines', *Journal of Experimental Psychology: General*, **149**(8): 1608.

Bak-Coleman, J. B., I. Kennedy, M. Wack, A. Beers, J. S. Schafer, E. S. Spiro, K. Starbird and J. D. West (2022), 'Combining interventions to reduce the spread of viral misinformation', *Nature Human Behaviour*, **6**: 1372–1380.

Benkelman, S. and H. Mantas (2020), *Can an Accuracy 'Nudge' Help Prevent People from Sharing Misinformation?* Poynter. https://www.poynter.org/fact-checking/2020/can-an-accuracy-nudge-help-prevent-people-from-sharing-misinformation/

Berinsky, A. J., M. F. Margolis and M. W. Sances (2014), 'Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys', *American Journal of Political Science*, **58**(3): 739–753.

Bode, L. and E. Vraga (2021), 'Value for correction: Documenting perceptions about peer correction of misinformation on social media in the context of Covid-19', *Journal of Quantitative Description: Digital Media*, **1**. https://doi.org/10.51685/jqd.2021.016

Brashier, N. M., G. Pennycook, A. J. Berinsky and D. G. Rand (2021), 'Timing matters when correcting fake news', *Proceedings of the National Academy of Sciences*, **118**(5): e2020043118.

Broniatowski, D. A. and V. F. Reyna (2018), 'A formal model of fuzzy-trace theory: variations on framing effects and the allais paradox', *Decision*, **5**(4): 205–252.

Broniatowski, D. A. and V. F. Reyna (2020), 'To illuminate and motivate: a fuzzy-trace model of the spread of information online', *Computational and Mathematical Organization Theory*, **26**(4): 431–464.

Chater, N. and G. Loewenstein (2023), 'The i-frame and the s-frame: how focusing on individual-level solutions has led behavioral public policy astray', *Behavioral and Brain Sciences*, **46**: e147.

Cialdini, R. (1984), *Influence: The Psychology of Persuasion*. New York, NY: Harper Collins.

Coppock, A., K. Gross, E. Porter, E. Thorson and T. J. Wood (2023), 'Conceptual replication of four key findings about factual corrections and misinformation during the 2020 US election: evidence from panel-survey experiments', *British Journal of Political Science*, **53**(4): 1328–1341.

Darley, J. M. and B. Latane (1968), 'Bystander intervention in emergencies: diffusion of responsibility', *Journal of Personality and Social Psychology*, **8**(4, Pt.1): 377–383.

DellaVigna, S. and E. Linos (2022), 'RCTs to scale: comprehensive evidence from two nudge units', *Econometrica*, **90**(1): 81–116.

Druckman, J. N. and T. J. Leeper (2012), 'Learning more from political communication experiments: pretreatment and its effects', *American Journal of Political Science*, **56**(4): 875–896.

Fischer, P., J. I. Krueger, T. Greitemeyer, C. Vogrincic, A. Kastenmüller, D. Frey, M. Heene, M. Wicher and M. Kainbacher (2011), 'The bystander-effect: a meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies', *Psychological Bulletin*, **137**(4): 517–537.

Gavin, L., J. McChesney, A. Tong, J. Sherlock, L. Foster and S. Tomsa (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*. https://doi.org/10.1037/tmb0000086

Gelman, A. (2022), The real problem of that nudge meta-analysis is not that it includes 12 papers by noted fraudsters: it's the Gigo of it all. https://statmodeling.stat.columbia.edu/2022/01/10/the-real-problem-of-that-nudge-meta-analysis-is-not-that-it-include-12-papers-by-noted-fraudsters-its-the-gigo-of-it-all/

Gottlieb, J. and C. S. Carver (1980), 'Anticipation of future interaction and the bystander effect', *Journal of Experimental Social Psychology*, **16**(3): 253–260.

Graham, M. H. (2023), 'Measuring misperceptions?', *American Political Science Review*, **117**(1): 80–102.

Greene, C. M. and G. Murphy (2021), 'Quantifying the effects of fake news on behavior: evidence from a study of Covid-19 misinformation', *Journal of Experimental Psychology: Applied*, **27**(4): 773.

Guess, A., J. Nagler and J. Tucker (2019), 'Less than you think: prevalence and predictors of fake news dissemination on facebook', *Science Advances*, **5**(1): eaau4586.

Guess, A., B. Nyhan and J. Reifler (2020a), 'Exposure to untrustworthy websites in the 2016 US election', *Nature Human Behaviour*, **4**(5): 472–480.

Guess, A. M., M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler and N. Sircar (2020b), 'A digital media literacy intervention increases discernment between mainstream and false news in the United States and India', *Proceedings of the National Academy of Sciences*, **117**(27): 15536–15545.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Hoes, E., B. Aitken, J. Zhang, T. Gackowski and M. Wojcieszak (2024), 'Prominent misinformation interventions reduce misperceptions but increase scepticism', *Nature Human Behaviour*, 1–9.

Howe, L. C. and J. A. Krosnick (2017), 'Attitude strength', *Annual Review of Psychology*, **68**(1): 327–351. PMID: 27618943.

Irving, D., R. W. A. Clark, S. Lewandowsky and P. J. Allen (2022), 'Correcting statistical misinformation about scientific findings in the media: causation versus correlation', *Journal of Experimental Psychology: Applied*, **28**(1): 1–9.

Jahanbakhsh, F., A. X. Zhang, A. J. Berinsky, G. Pennycook, D. G. Rand and D. R. Karger (2021), 'Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media', *Proceedings of the ACM on Human–Computer Interaction*, **5**(CSCW1): 1–42.

Jigsaw (2023), Prebunking to build defenses against online manipulation tactics in Germany. https://medium.com/jigsaw/prebunking-to-build-defenses-against-online-manipulation-tactics-in-germany-a1dbfbc67a1a

Kuklinski, J. H., P. J. Quirk, J. Jerit, D. Schwieder and R. Rich (2000), 'Misinformation and the currency of democratic citizenship', *Journal of Politics*, **62**(3): 790–816.

Larson, H. J. and D. A. Broniatowski (2021), 'Why debunking misinformation is not enough to change people's minds about vaccines', *American Journal of Public Health*, **111**(6).

Lin, H., H. Garro, N. Wernerfelt, J. Shore, A. Hughes, D. Deisenroth, N. Barr, A. Berinsky, D. Eckles, G. Pennycook and D. Rand (2024), Reducing misinformation sharing at scale using digital accuracy prompt ads. https://doi.org/10.31234/osf.io/u8anb

Madrian, B. C. and D. F. Shea (2001), 'The power of suggestion: inertia in 401 (k) participation and savings behavior', *The Quarterly Journal of Economics*, **116**(4): 1149–1187.

Maier, M., F. Bartoš, T. Stanley, D. R. Shanks, A. J. Harris and E.-J. Wagenmakers (2022), 'No evidence for nudging after adjusting for publication bias', *Proceedings of the National Academy of Sciences*, **119**(31): e2200300119.

Martel, C., S. Rathje, C. J. Clark, G. Pennycook, J. J. Van Bavel, D. G. Rand and S. van der Linden (2024), 'On the efficacy of accuracy prompts across partisan lines: an adversarial collaboration', *Psychological Science*, **35**(4): 435–450.

Mertens, S., M. Herberz, U. J. J. Hahnel and T. Brosch (2022), 'The effectiveness of nudging: a meta-analysis of choice architecture interventions across behavioral domains', *Proceedings of the National Academy of Sciences*, **119**(1): e2107346118.

Milkman, K. L., L. Gandhi, M. S. Patel, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, J. Rothschild, J. E. Bogard, I. Brody, C. F. Chabris, E. Chang, G. B. Chapman, J. E. Dannals, N. J. Goldstein, A. Goren, H. Hershfield, A. Hirsch, J. Hmurovic, S. Horn, D. S. Karlan, A. S. Kristal, C. Lamberton, M. N. Meyer, A. H. Oakes, M. E. Schweitzer, M. Shermohammed, J. Talloen, C. Warren, A. Whillans, K. N. Yadav, J. J. Zlatev, R. Berman, C. N. Evans, R. Ladhania, J. Ludwig, N. Mazar, S. Mullainathan, C. K. Snider, J. Spiess, E. Tsukayama, L. Ungar, C. Van den Bulte, K. G. Volpp and A. L. Duckworth (2022), 'A 680,000-person megastudy of nudges to encourage vaccination in pharmacies', *Proceedings of the National Academy of Sciences*, **119**(6): e2115126119.

Mummolo, J. and E. Peterson (2019), 'Demand effects in survey experiments: an empirical assessment', *American Political Science Review*, **113**(2): 517–529.

Nyhan, B. (2020), 'Facts and myths about misperceptions', *Journal of Economic Perspectives*, **34**(3): 220–236.

Nyhan, B., E. Porter, J. Reifler and T. Wood (2019), 'Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability', *Political Behavior*, **42**: 939–960.

Orchinik, R., C. Martel, D. G. Rand and R. Bhui (2023), Uncommon errors: adaptive intuitions in high-quality media environments increase susceptibility to misinformation. *PsyArXiv*.

Oreopoulos, P. and U. Petronijevic (2019), 'The remarkable unresponsiveness of college students to nudging and what we can learn from it', *No. w26059. National Bureau of Economic Research*.

Pennycook, G. and D. G. Rand (2019), 'Fighting misinformation on social media using crowd-sourced judgments of news source quality', *Proceedings of the National Academy of Sciences*, **116**(7): 2521–2526.

Pennycook, G. and D. G. Rand (2022a), 'Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation', *Nature Communications*, **13**(1): 2333.

Pennycook, G. and D. G. Rand (2022b), 'Nudging social media toward accuracy', *The ANNALS of the American Academy of Political and Social Science*, **700**(1): 152–164. PMID: 35558818.

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press

Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu and D. G. Rand (2020), 'Fighting Covid-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention', *Psychological Science*, **31**(7): 770–780. PMID: 32603243.

Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles and D. G. Rand (2021), 'Shifting attention to accuracy can reduce misinformation online', *Nature*, **592**(7855): 590–595.

Philpot, R., L. S. Liebst, M. Levine, W. Bernasco and M. R. Lindegaard (2020), 'Would i be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts', *American Psychologist*, **75**(1): 66.

Porter, E. and Y. R. Velez (2021), 'Placebo selection in survey experiments: an agnostic approach', *Political Analysis*, 1–14.

Porter, E. and T. J. Wood (2021), 'The global effectiveness of fact-checking: evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom', *Proceedings of the National Academy of Sciences*, **118**(37): e2104235118.

Porter, E. and T. J. Wood (2024), 'Factual corrections: concerns and current evidence', *Current Opinion in Psychology*, **55**: 101715.

Pretus, C., C. Servin-Barthet, E. A. Harris, W. J. Brady, O. Vilarroya and J. J. Van Bavel (2023), 'The role of political devotion in sharing partisan misinformation and resistance to fact-checking', *Journal of Experimental Psychology: General*, **152**(11): 3116–3134.

Prike, T., P. Blackley, B. Swire-Thompson and U. K. Ecker (2023), 'Examining the replicability of backfire effects after standalone corrections', *Cognitive Research: Principles and Implications*, **8**(1): 39.

Reyna, V. F., D. A. Broniatowski and S. M. Edelson (2021), 'Viruses, vaccines, and Covid-19: explaining and improving risky decision-making', *Journal of Applied Research in Memory and Cognition*, **10**(4): 491–509.

Rogers, T., N. J. Goldstein and C. R. Fox (2018), 'Social mobilization', *Annual Review of Psychology*, **69**: 357–381.

Roozenbeek, J., A. L. J. Freeman and S. van der Linden (2021), 'How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020)', *Psychological Science*, **32**(7): 1169–1178. PMID: 34114521.

Roozenbeek, J., S. Van Der Linden, B. Goldberg, S. Rathje and S. Lewandowsky (2022), 'Psychological inoculation improves resilience against misinformation on social media', *Science Advances*, **8**(34): eabo6254.

Rovira, A., R. Southern, D. Swapp, C. Campbell, J. J. Zhang, M. Levine and M. Slater (2021), 'Bystander affiliation influences intervention behavior: a virtual reality study', *SAGE Open*, **11**(3): 21582440211040076.

Sunstein, C. R. (2013), 'Nudges vs. shoves', *Harvard Law Review Forum*, **127**: 210.

Sunstein, C. R. (2017), 'Nudges that fail', *Behavioural Public Policy*, **1**(1): 4–25.

Thaler, R. and C. Sunstein (2008), *Nudge: Improving Decisions about Health, Wealth and Happiness*. New Haven, CT: Yale University Press.

van Bommel, M., J.-W. van Prooijen, H. Elffers and P. A. Van Lange (2012), 'Be aware to care: public self-awareness leads to a reversal of the bystander effect', *Journal of Experimental Social Psychology*, **48**(4): 926–930.

Vraga, E. K., L. Bode and M. Tully (2021), 'Creating news literacy messages to enhance expert corrections of misinformation on twitter', *Communication Research*, 0093650219898094.

Zhang, B., M. Mildenberger, P. D. Howe, J. Marlon, S. A. Rosenthal and A. Leiserowitz (2020), 'Quota sampling using facebook advertisements', *Political Science Research and Methods*, **8**(3): 558–564.

Cite this article: Porter, E., T.J. Wood, D.A. Broniatowski and P. Hosseini (2026), 'Shoves, nudges and combating misinformation: evidence on a new approach', *Behavioural Public Policy* **10**(3): 558–574. https:// doi.org/10.1017/bpp.2024.42

https://doi.org/10.1017/bpp.2024.42 Published online by Cambridge University Press