Journal of Public Economics 245 (2025) 105345

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Debunking "fake news" on social media: Immediate and short-term effects of fact-checking and media literacy interventions☆

Lara Marie Berger [a,*], Anna Kerkhof [b,c], Felix Mindl [a,d], Johannes Münster [a]

[a] University of Cologne, Germany
[b] ifo Institute for Economic Research, Germany
[c] CESifo, Germany
[d] iwp Institute for Economic Policy, Germany

ARTICLE INFO

JEL classification:
L51
L82
Z18

Keywords:
Covid
Environment
Facebook
Fact-checking
Fake news
Media literacy
Misinformation
Nutrition
Social media
Supplements
Survey experiment
Vaccine

ABSTRACT

We conduct a randomized survey experiment to compare the immediate and short-term effects of fact-checking to a brief media literacy intervention. We show that fact-checking primarily affects the specific fake news it directly addresses, whereas media literacy helps to distinguish between false and correct information more generally, both immediately and around two weeks after the intervention. A plausible mechanism is that media literacy enables participants to critically evaluate social media postings, while fact-checking fails to enhance their skills as much. Our results promote media literacy as an effective tool to fight fake news, that is cheap, scalable, and easy-to-implement.

1. Introduction

The emergence and spread of "fake news" – i.e., false or misleading information presented as news – has led to widespread concerns (e.g., Lazer et al., 2018). Social media like Facebook and X (formerly Twitter) are especially prone catalysts for the evolution of fake news and have consequently come to the fore of public and academic debates. Indeed, recent evidence suggests that 50% of users who see fake news on social media say that they believe them (Allcott and Gentzkow, 2017).

What helps users to distinguish between false and correct information on social media? Policymakers support fact-checkers on the one hand, and media literacy initiatives on the other.[1] Independent fact-checking organizations complement such campaigns, and large social

☆ We thank the Co-Editor, Maria Petrova, and three anonymous referees, Maja Adena, Kai Barron, Anna Bindler, Esther Blanco, Grazia Cecere, Felix Chopra, Oliver Falck, Marcel Garz, Lorenz Gschwent, Julian Harke, Paul Hufe, Jonas Loebbing, Nikola Noske, Arianna Ornaghi, Dominik Rehse, Heiner Schumacher, Andreas Steinmayr, Christian Traxler, Maiting Zhuang and participants of various seminars and conferences for helpful comments and suggestions. Olga Andreeva, Lara Mai and Aleksander Vasic provided excellent research assistance. Anna Kerkhof acknowledges financial support from the Joachim Herz Stiftung, Germany and by the Bavarian State Ministry of Science and the Arts in the framework of the bidt Graduate Center for Postdocs. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy -EXC 2126/1-390838866. The experiment was approved by the Ethics Committee of the University of Cologne (reference 210014JM) and pre-registered in the AEA Registry under registry number AEARCTR-0008199. All errors are our own.

* Corresponding author.
E-mail addresses: lara.berger@wiso.uni-koeln.de (L.M. Berger), kerkhof@ifo.de (A. Kerkhof), mindl@wiso.uni-koeln.de (F. Mindl), johannes.muenster@uni-koeln.de (J. Münster).

[1] See, e.g., https://digital-strategy.ec.europa.eu/en/library/disinformation-threat-democracy-brochure and https://digital-strategy.ec.europa.eu/en/policies/online-disinformation (viewed August 2022) for further information on efforts by the European Union.

[2] Guriev et al. (2023) compare the effect of an attention nudge (which could be thought of as a minimalistic media literacy enhancement) to fact-checking and other policies with sharing of fake posts as an outcome.

https://doi.org/10.1016/j.jpubeco.2025.105345
Received 4 September 2023; Received in revised form 14 February 2025; Accepted 25 February 2025
Available online 5 April 2025

media have started to implement AI-based procedures that flag suspicious content, too (Allen et al., 2021). Yet, it is unclear whether these remedies function as desired: empirical evidence on the effectiveness of fact-checking is mixed (Vraga and Bode, 2017; Jerit and Zhao, 2020), knowledge on the impact of media literacy is scarce (Guess et al., 2020), and a direct comparison of the effect of these interventions on the beliefs and attitudes of consumers does not exist at all.[2]

We address this gap with a large-scale randomized survey experiment on the immediate and short-term effects of fact-checking and media literacy interventions. In the experiment, we expose a sample of German residents to false and correct statements on health-related topics – Corona vaccines and nutrition – that we retrieve from the German version of Facebook ("fakes" and "facts").[3] One group of participants receives additional fact-checks that debunk some of the fakes explicitly. Another group gets ten "Tips to spot fake news" before exposure to the fakes and facts as a brief media literacy intervention. Then, we compare the two treatment groups to participants who do not receive an intervention, which informs us about the immediate impact of our interventions. To study the interventions' efficacy in the short-run, we survey the same participants around two weeks later in a second, similar experimental wave without any additional interventions.

Our results demonstrate that the effectiveness of fact-checking tends to be limited to the fakes that are being corrected, whereas media literacy helps to distinguish between fakes and facts more generally, both immediately and about two weeks after the intervention. A plausible explanation is that the media literacy intervention raises participants' attention and enables them to critically evaluate the postings' accuracy. Fact-checking, in contrast, turns participants into passive recipients of the specific corrections and thus fails to markedly enhance their skills.[4]

Specifically, we consider three main outcomes: the perceived credibility of fakes and facts, factual knowledge on the topics discussed therein, and attitudes towards Corona vaccination and dietary supplements (the fakes on nutrition promote the consumption of needless protein and vitamin preparations). The idea is to study a coherent cognitive chain: Do the interventions reduce the perceived credibility of fakes (but not of facts)? If yes, does that translate into better factual knowledge? If yes, does this entail a change in attitudes?

We find that both interventions reduce the credibility of fakes on Corona vaccines (which are corrected by fact-checks) immediately, but only the media literacy intervention reduces the credibility of fakes on nutrition (which are not corrected by fact-checks), both immediately and in the short-run (two weeks later). Moreover, both interventions improve participants' factual knowledge immediately, but only the media literacy intervention in the short-run. Finally, while the media literacy intervention raises participants' willingness to get vaccinated (or boostered) against Covid-19 both immediately and in the short-run, fact-checking has no such effect. Crucially, neither intervention reduces the credibility of facts or factual knowledge on the topics discussed therein, i.e., participants do not become more skeptical towards social media postings per se. Hence, in an environment where not every posting can be fact-checked, media literacy interventions are likely to be more effective than fact-checking on average.

Our subgroup analyses reveal that participants who are well informed from the beginning are less likely to benefit from the interventions than participants whose prior beliefs are further away from the truth. In particular, both the fact-checking and the media literacy intervention are *more* effective for fakes on Corona vaccines for supporters of the German AfD ("Alternative for Germany"), a far-right populist party known for spreading misinformation on Covid-19. For fakes on nutrition, where participants' beliefs are much more alike, we observe no such effect. Computing persuasion rates à la DellaVigna and Kaplan (2007) shows that this result can only be partly explained by differences in the proportion of participants who are left to be convinced. However, we also provide evidence that AfD supporters are less certain about their prior knowledge than non-AfD supporters, so the former group may also be easier to convince. In contrast to that, we do not find any systematic effect heterogeneity in terms of education, age, social media usage, support of Corona policy measures, or prior knowledge on current events, health, and nutrition.

A plausible mechanism for our main results is that the media literacy intervention raises participants' attention and enables them to critically evaluate the Facebook postings, while fact-checking fails to markedly enhance their skills. To support the plausibility of this explanation, we show that participants who receive the media literacy intervention are more likely to actively search for further information when they respond to our questions than participants who receive the fact-checking or no intervention at all. Also, media literacy helps participants to better consider untrustworthy elements in fakes and trustworthy elements in facts. Fact-checking, in contrast, has no such effect. Moreover, we show that participants in a follow-up experiment who received a media literacy intervention spend more time reading the posts than those without any intervention, while participants who received fact-checks spend less time reading.[5]

While our main analysis illustrates the effectiveness of fact-checking and media literacy interventions in an environment where all participants see fakes and facts, it is agnostic about the extent to which the interventions are able to reverse the harm the fakes are causing. To better interpret the magnitude of our coefficients in that regard, we also compare the three main treatment groups to participants who do not see any Facebook postings at all. We find that exposure to fake news substantially impairs participants' factual knowledge, and that neither the fact-checking nor the media literacy intervention can fully offset the effect. Participants' attitudes on Corona vaccination and dietary supplements, in contrast, are hardly affected by fakes and especially the media literacy intervention can effectively repeal that impact.

To better assess the external validity of our findings, we conducted a follow-up study, applying the interventions to a new context. Specifically, we exposed participants to misinformation about environmental topics instead of Covid-19.[6] This additional experiment demonstrates that both fact-checking and media literacy interventions can be effective in different contexts and at different times. Moreover, our follow-up study indicates that the estimated effectiveness of the media literacy intervention is significantly greater than that of fact-checking, reinforcing our conclusion that media literacy has a broader impact than fact-checks.

As far as we know, we are the first who pursue a clean comparison of fact-checking and media literacy interventions as a means to diminish the belief in fake news, whereby we provide a valuable contribution to public and academic debates. Since public resources to combat fake news are limited, it is of utmost importance to understand when and why which remedies are most effective, so that time, money, and effort can be efficiently allocated. Pennycook and Rand (2021), for instance,

[3] Online misinformation about health can have severe consequences for quality of life and even for mortality risks (see Swire-Thompson and Lazer (2020) for a review of the literature on online misinformation about health, and Allen et al. (2024) for a study of vaccine related misinformation on Facebook).

[4] While we do not find any evidence for an effect of fact-checking on the perceived credibility of non-fact-checked posts in our main experiment, a follow-up study reveals a small, but statistically significant effect. For details see Appendix E.

[5] We did not collect the exact time spent with each post in the main experiment.

[6] Gundersen et al. (2022) and Erbaugh et al. (2024) provide reviews about online misinformation concerning the environment. For example, a lot of misinformation about climate science and climate change circulates on social media, delaying climate action (Erbaugh et al., 2024; Lewandowsky, 2021; Gundersen et al., 2022).

stress that professional fact-checking is "simply not scalable" (p.396), as it requires substantial time and effort to examine a particular claim, and even if the claim is eventually tagged as false, the warning is likely to be missing during the peak of its spread. Similar caveats apply to the recent approach of (human) crowd sourced fact-checking (Allen et al., 2021), which is furthermore limited to the linguistic proficiency of crowd workers. AI-based procedures are more efficient than human search, but still in their infancy. Specifically, while they are able to detect potential misinformation, the fact-checking as such still requires human assessment.[7] We show that in an environment where only a small proportion of fake news can ever be fact-checked, a brief media literacy intervention is likely to be more effective than fact-checking on average. Moreover, given that displaying a small number of tips and heuristics to users of social media is cheap, scalable, and easy-to-implement, our results promote media literacy interventions as a (potentially more) powerful tool to combat fake news.

Our paper advances the surprisingly small body of research on (digital) media literacy as a means to fight fake news (Guess et al., 2020; Roozenbeek et al., 2022) and adds to a recent literature that acknowledges the limits of fact-checking (see Jerit and Zhao, 2020, for a review). E.g., Pennycook and Rand (2019) argue that many users fall for fake news because they fail to reflect; similarly, Pennycook et al. (2020, 2021) show that users share false claims partly because they do not think sufficiently about whether or not the content is accurate. Guriev et al. (2023) show that an attention nudge (which could be thought of as a minimalistic media literacy enhancement) can reduce sharing of fake content on X to a greater extent than fact-checks. Consistent with what we find, such results advocate media literacy interventions that help users to critically evaluate social media postings as a promising avenue, while assorting fact-checking – which fails to markedly enhance users' skills – as less effective.

The remainder of the paper is organized as follows. Section 2 reviews the related literature on social media, misinformation, and education interventions. Section 3 illustrates the experimental setup and implementation; moreover, we describe our empirical strategy. Section 4 presents our main results, where we compare the effectiveness of fact-checking and media literacy on the credibility of and factual knowledge on fakes, as well as on participants' attitudes. In Section 5, we show that an increase in attention and the ability to critically evaluate social media postings on behalf of the media literacy intervention is a plausible mechanism for our results. Section 6 presents further results and robustness checks, Section 7 concludes. We describe our follow-up experiment which includes stimuli on environmental topics in Appendix E.

2. Related literature

Social media and UGC. Our paper is related to two strands of literature. First, it adds to the vibrant and interdisciplinary research on social media and user-generated content (reviewed by Luca, 2015; Zhuravskaya et al., 2020), where it is particularly close to analyses of fake news. This subfield can be further divided into studies on the emergence and spread of fake news (e.g., Allcott and Gentzkow, 2017; Lazer et al., 2018; Guess et al., 2018, 2019; Grinberg et al., 2019; Vosoughi et al., 2018; Bursztyn et al., 2023), and inquiries of potential remedies (reviewed by Lewandowsky et al., 2012; Jerit and Zhao, 2020; Allen et al., 2021).[8] The latter literature focuses on corrective interventions like fact-checking: While Bode and Vraga (2015), Vraga and Bode (2017), and Henry et al. (2022), among others, support its effectiveness, other papers find no or even "backfire" effects (e.g., Nyhan and Reifler, 2010, 2015), or they document mixed results, whereby fact-checking improves users' factual knowledge, but struggles to change more deep-rooted perceptions and attitudes (Barrera et al., 2020; Nyhan et al., 2020). Indeed, the efficacy of fact-checking varies significantly depending on the outcome being considered. The consensus in the literature suggests that while fact-checking has limited effect on correcting fundamental beliefs, it is effective in influencing actions, such as reducing the sharing rate of false news. Ex post fact-checking, in particular, may be ineffective, as it is challenging to correct beliefs after exposure to false or misleading information (Swire et al., 2017; Nyhan et al., 2020; Barrera et al., 2020). Conversely, Henry et al. (2022) demonstrate that both imposed and voluntary fact-checking reduce the sharing of false statements by over 25%. Recent work by Drolsbach et al. (2024) extends this discussion by examining X's community-based fact-checking system "Community Notes". Their findings highlight that explanatory context provided by community notes not only fosters trust in fact-checks across political divides, but also enhances users' ability to discern misleading content.

The literature on fact-checking varies in terms of whether corrections are provided before or after exposure to false information. Both approaches appear to be somewhat effective, but there is no clear consensus on which is superior (Swire-Thompson et al., 2021). For instance, Lewandowsky et al. (2012), Ecker et al. (2022), and Lewandowsky and Van Der Linden (2021) advocate for *pre-bunking*, where corrective information is presented before exposure to fakes. They argue this method more effectively prevents the "continued influence effect" (Lewandowsky et al., 2012), where false information lingers in memory once learned. However, other studies find that the order of presenting fakes and corrections does not significantly impact effectiveness (Swire-Thompson et al., 2021), or that corrections shown after the fakes have a greater effect (Brashier et al., 2021; Dai et al., 2021). In our study, we align with Facebook's current practice by presenting corrections before showing the fake information to participants.[9]

A novel line of research examines the efficacy of removing misleading content or platform users altogether. E.g., Broniatowski et al. (2023) and Gu et al. (2022) evaluate the impact of Facebook's strict misinformation policy in March 2019 on user endorsements of vaccine content. Similarly, Mitts et al. (2022) and Ali et al. (2021) evaluate the effectiveness of banning prominent groups and figures who spread false or misleading content on Facebook and X. These studies agree that removing content or users is not very effective, as it does not reduce overall engagement with misleading content, and arguably increases the production of fake news elsewhere. In contrast, Ershov and Morales (2024) evaluate a user-centric intervention where Twitter introduced friction to content sharing during the 2020 U.S. presidential election.

Studies on alternative ways to combat fake news are relatively rare. One notable exception is Guess et al. (2020), who assess the effectiveness of Facebook's "Tips to Spot False News" on discernment between mainstream and false news headlines both among a nationally representative sample in the US and a highly educated online sample in India. Relatedly, Roozenbeek et al. (2022) use five short videos that inoculate people against manipulation techniques commonly used in misinformation and find that they improve manipulation technique recognition, boost confidence in spotting these techniques, increase users' ability for truth discernment as well as the quality of their sharing decisions. Bak-Coleman et al. (2022) derive a generative model of fake news engagement and use Bayesian simulation techniques to show that

[7] See EU Horizon, URL: https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news (viewed August 2023).

[8] Because of the topics of the statements we use as experimental stimuli, our main experiment is particularly related to the literature on misinformation about health (see Swire-Thompson and Lazer (2020) for a review), and our follow-up study to the literature on misinformation about the environment (see Gundersen et al. (2022) and Erbaugh et al. (2024) for reviews).

[9] In a follow-up experiment, we randomize the order in which participants view the fake and the corresponding fact-check. For more details see Appendix E.3.4.

fact-checking, removal of misleading content and users, and nudges towards accuracy are efficient in combination, but unlikely to work in isolation. Guriev et al. (2023) compare the effect of an attention nudge (which could be thought of as a minimalistic media literacy enhancement) to fact-checking and show that it is more effective than fact-checking and other policies in reducing the sharing rates of fake posts on X.

We contribute to this literature in several ways. First, we pursue a clean comparison of the effect of fact-checking and media literacy interventions on the belief of consumers, which has not been done so far. In particular, our experimental setup allows us to study the immediate and short-term effects of fact-checking and media literacy interventions in one and the same environment, whereby we can observe when and why which remedy is most effective. In addition, we provide evidence for potential mechanisms behind our results, shifting the research focus from asking *whether* fact-checking and media literacy interventions are effective tools to fight fake news to studying *how* they work and *in which case* they fail or succeed.

Very closely related to our study are Barrera et al. (2020) and Guess et al. (2020). Barrera et al. (2020) use a randomized online experiment to expose voters to fakes, facts, and fact-checks on immigration in France. Participants are then asked about their posterior beliefs on topics related to immigration, their opinions on immigration policy, as well as their voting intentions. Similar to what we find, Barrera et al. (2020) demonstrate that fake news are highly persuasive, and while fact-checking enhances factual knowledge, it fails to offset the fakes' effect on voting intentions.[10] Guess et al. (2020) examine the impact of a digital media literacy intervention on the perceived accuracy of false and correct news headlines and show that participants' ability for truth discernment increases.

Our results largely confirm these findings, but we extend the preceding analyses in several ways. First, we explore the immediate *and* short-term effects of fact-checking *and* media literacy interventions on fakes *and* facts within one experiment, which allows us to directly compare these remedies and draw a sophisticated picture of how and when which type of intervention works. Likewise, we consider a broad range of coherent outcomes – credibility, factual knowledge, and attitudes – and complement our analysis with a thorough examination of potential mechanisms. Finally, we use postings from social media that actually exist and whose content is not necessarily politically loaded, demonstrating that our results hold beyond the partisan context.

We also contribute to a growing body of research arguing that users fall for fake news because they fail to pay sufficient attention (e.g., Pennycook and Rand, 2019; Loewenstein and Wojtowicz, 2023). Pennycook et al. (2020, 2021), for instance, show that users frequently share misinformation because they do not focus on accuracy; politically motivated reasoning, in contrast, seems to play a minor role. Our results are in line with such findings, because we demonstrate that media literacy interventions – which raise users' attention and help them to actively distinguish between fakes and facts – are on average more effective than just passively receiving fact-checks. Moreover, in contrast to previous findings on motivated reasoning (e.g., Lewandowsky et al., 2012; Jerit and Zhao, 2020), we find that our interventions are more effective for supporters of a far-right political party, who are initially much more likely to oppose Corona vaccination. This, too, is consistent with the above line of thought, whereby it is often a lack of attention rather than partisanship why people fall for fake news.

Education interventions. Second, our paper is related to the literature on education interventions, most of which focuses on financial literacy. In accordance with our results on the media literacy intervention, a meta-analysis of 76 randomized experiments by Kaiser et al. (2022) reveals that financial education interventions have, on average, positive causal treatment effects on financial knowledge and behavior. The treatment effects are economically meaningful in size and comparable to those realized by education interventions in math and reading (e.g., Hill et al., 2008; Cheung and Slavin, 2016; Fryer Jr., 2017), health (e.g., Rooney and Murray, 1996; Noar et al., 2007), and energy saving behavior (e.g., Karlin et al., 2015). In particular, Kaiser et al. (2022) report that the average effect size of financial education interventions corresponds to about 0.123 standard deviation units. This is only about half as much as what we see for the impact of our media literacy intervention on the credibility of fakes, and roughly as much as what we observe for the impact on factual knowledge on average. However, Kaiser et al. (2022) stress that the effectiveness of education interventions diminishes over time. Given that the average effect size from their meta-analysis also includes interventions that were evaluated after several weeks or months, it is not surprising that the effect sizes from our media literacy intervention are comparably large.

A plausible explanation for the larger effectiveness of the media literacy relative to the fact-checking intervention is that the former raises participants' attention and enables them to critically evaluate the postings' accuracy, whereas the latter only helps them to update beliefs about one specific fact and fails to markedly enhance their more general skills. This matches the findings by Kaiser and Menkhoff (2022), who study different types of interventions offered to small-scale retailers in Uganda and show that "active learning" has a positive effect on savings and investment outcomes while traditional lecturing is ineffective. Comparable results on the advantages of active as opposed to passive learning have been documented in other domains, including science, technology, engineering, and maths (e.g., Deslauriers et al., 2011; Ruiz-Primo et al., 2011; Freeman et al., 2014). Relatedly, Drexler et al. (2014) show that a heuristics-based approach relying on "rules-of-thumb-training" – such as our ten tips to spot false news – generates larger behavioral impacts than the teaching of full curricula.

3. Experimental design

3.1. Survey flow

We start by randomizing the participants of our online survey experiment into one out of five groups of approximately equal size: (i) NoIntervention, (ii) FactChecking, (iii) MediaLiteracy, (iv) JustFacts, and (v) PassiveControl. To study both the immediate and short-term effects of our interventions, we conduct two waves of the experiment, where we re-invite the same participants one week after they completed Wave I and allocate them to the same treatment group as before. Fig. 1 gives an overview of our survey flow, Appendix F provides the original survey flow as extracted from our survey software Qualtrics, further details are discussed below.[11]

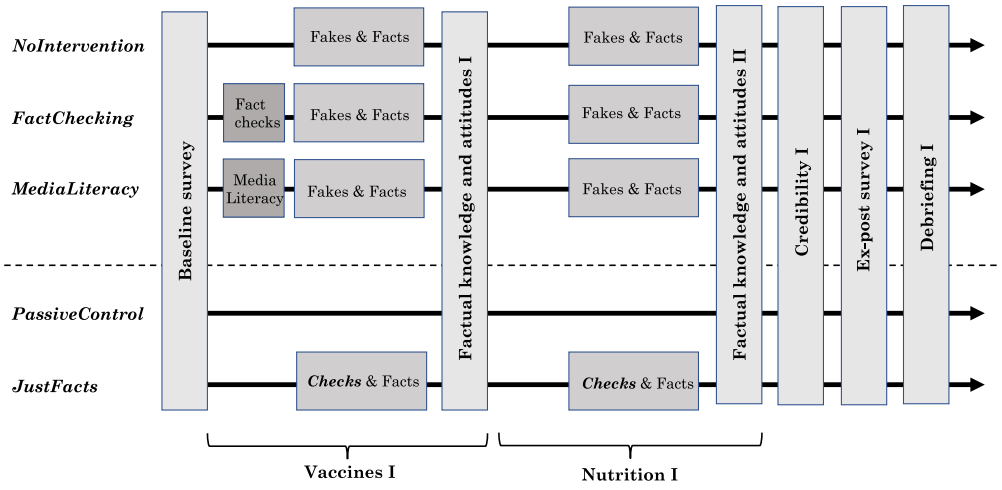
3.1.1. Wave I

Baseline survey. All participants start with a baseline survey on standard demographics such as age, gender, family status, household income, education, profession, and personality traits ("big five"). In addition, we inquire participants' prior knowledge on current events, health, and nutrition.[12] To avoid priming effects on subsequent questions, we do not include questions about the figures of interest after the interventions, but ask (i) how many days Joe Biden has been President of the United States, (ii) when to see a doctor in case of high temperature, and (iii) how many servings of fruit and vegetables are officially

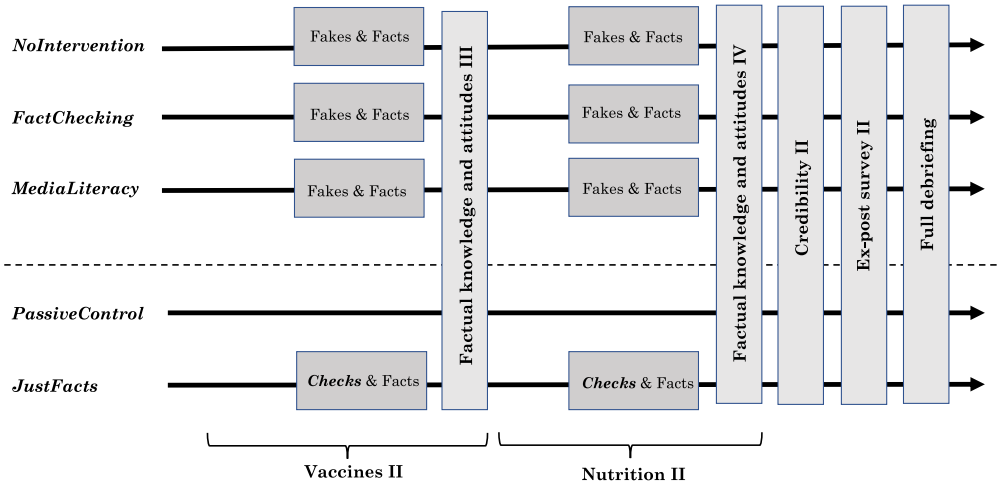
[10] Similar results are presented by Nyhan et al. (2020).

[11] We preregistered our two interventions and five treatment groups as described above. We sketched the survey flow but did not preregister it in full detail. We also did not preregister which fakes, facts, and fact-checks we would display exactly. Please see Appendix B for details.

[12] We preregistered the full set of control variables. Please see Appendix B for details.



(a) Overview Wave I



(b) Overview Wave II

Fig. 1. Survey flow.

recommended per day. To measure the strength of participants' prior beliefs, we also ask how certain they are about the accuracy of their responses on a 5-point Likert scale ranging from *Very uncertain* to *Very certain*. To maintain participants' engagement, we incorporated two attention checks: one placed in the middle of the baseline survey and another at the end (see Appendix F).

Vaccines. Participants in the NoIntervention, the FactChecking, and the MediaLiteracy group are shown two pieces of "fake news" ("fakes" henceforth) and two facts on Corona vaccines in randomized order.[13] The fakes and facts were manually collected from Facebook, i.e., we use screenshots of Facebook postings that actually exist. All fakes and facts were collected between March and May 2021.

We used two search strategies to collect the postings. First, we used Facebook's search function with German keywords equivalent to "covid", "corona", "vaccine" and "side effect". We then screened through all publicly available search results, including postings from individual profile pages and Facebook groups. Second, we started from the fake news repositories of renowned German fact-checking initiatives like *AFP* and *Correctiv.org*. Specifically, we skimmed through all registered fake news and checked which of them stem from Facebook and deal with the topic of interest. The collected postings from either search strategy had to meet the following criteria: They were only included in the experiment if we could find appropriate fact-checks debunking the false information, they had to be either about Covid-19 vaccines or dietary supplements and they had to be posted not too long ago. In addition, all fakes and facts had to contain a concrete numerical value (e.g., "50 people died after vaccination in a Sana clinic") that we could later on ask for. Using this catalogue of criteria led to the pool of posts we use. See Appendix A.3 for all fakes and facts that are part of the experiment.[14]

[13] We find no evidence for order effects: Our main results are unaffected when we control for the order of the fakes and facts. Moreover, we find no differences between participants who first saw a fake and participants who first saw a fact.

[14] To strengthen the external validity of our findings, a follow-up experiment exposes a new sample of participants to fakes and facts on environmental topics (see Appendix E.3.2).

Participants in the NoIntervention group do not receive further information. Participants in the FactChecking group, in contrast, receive additional fact-checks that explicitly debunk the false information (e.g., an official statement that the story about 50 deaths after vaccination in a Sana clinic is false). All fact-checks stem from sources that are commonly perceived as trustworthy (e.g., *Correctiv.org*). The fact-checks are shown *prior* to the fakes that they correct. We thereby follow the current procedure on Facebook, where false or misleading information – if detected – is overlain with a warning message that redirects the user to a fact-check; the original post can only be seen after the user closes the warning. Moreover, given that the timing of the intervention is known to be relevant (Lewandowsky et al., 2012; Brashier et al., 2021), displaying the fact-check prior to the respective fake makes the fact-checking better comparable to the media literacy intervention. See Appendix A.3 for all fact-checks that we use.[15]

Participants in the MediaLiteracy group receive Facebook's official "Tips to spot false news" *before* they are exposed to fakes and facts about Corona vaccines.[16] These tips actually exist on the platform and comprise ten short pieces of advice, including "Be skeptical of catchy headlines", "Look closely at the link", and "Investigate the source"; Appendix A.1 shows the full list. We inform our participants that these tips have been developed by Facebook itself. We display one tip per page and ask the participants to read them carefully before they proceed to the Facebook postings on Corona vaccines.

"Media literacy" is commonly understood as the ability to access, analyze, understand, and reflect on messages conveyed through mass media.[17] Our brief intervention, which offers ten tips to spot false news, cannot fully equip an individual with all the skills required for complete media literacy. However, these tips provide practical heuristics that enhance participants' awareness and aid in their critical evaluation of the fakes and facts that we present. Thus, while our "media literacy intervention" cannot fully develop media literacy, it reinforces existing skills and encourages participants to apply them effectively.

Note that participants in the NoIntervention group do not receive any placebo treatment such as short pieces of text with unrelated information (e.g., on financial education). First, going through the fact-checking or the media literacy intervention requires a relatively short amount of time, so we do not expect any differences in performance between the treatment groups owing to differences in fatigue. Second, we consider the scenario where participants read fakes and facts and nothing else as the most relevant and externally valid benchmark. In particular, we wish to examine if the fact-checking or the media literacy intervention help participants to better distinguish between fakes and facts relative to a situation where they receive no treatment at all, as would often be the case on social media platforms.

In contrast to the other groups, participants in the JustFacts and in the PassiveControl group are *not* exposed to fakes. While the PassiveControl group does not see any postings at all, participants in the JustFacts group receive the same two facts and fact-checks (without the corresponding fakes) as participants in the FactChecking group.[18] We can thereby infer our participants' average prior beliefs and attitudes from responses by the PassiveControl, and the impact of stand-alone fact-checks from the JustFacts group.

After exposure to the Facebook postings, we ask all participants four factual questions that are tailored to the fakes and facts just shown (see Appendix A.3 for all fakes and facts):

1. How many employees of Sana Kliniken died after being vaccinated against Corona in March 2021?
2. How much radioactive radiation does a dose of Corona vaccine emit (in microsievert)?
3. How effective is the Biontech/Pfizer vaccine in adolescents (in percentage)?
4. How many doses of the Russian Sputnik V vaccine has Bavaria ordered independently of the EU?

Each question asks for a specific number, and participants must give their answer through an input box, i.e., we do not provide a list of pre-defined options. To secure high quality answers, we use a bonus payment scheme that rewards participants whose answers are close to the true value.[19]

Next, we inquire all participants' willingness to get vaccinated against Covid-19. We start by asking each participant if he or she was already fully vaccinated by the time of the experiment. Based on their response, we partition the group into those who are fully vaccinated and those who are not. Then, we ask the former group about their willingness to get a booster injection as soon as it is officially recommended, and the latter about their willingness to get vaccinated against Covid-19 in general. Answers could be given on a 5-point Likert scale ranging from *Very likely* to *Very unlikely*. To avoid experimenter demand effects, we do not incentivize this question with a potential bonus payment (see Section 6.6.3 for further discussion). To maintain participants' engagement, we placed an attention check before proceeding to the second part of the survey (see Appendix F).[20]

Nutrition. The second part of Wave I is analogous to part one, except that we switch from Corona vaccines to nutritional topics, and that there are no further interventions (i.e., the setup is identical for participants in the NoIntervention, the FactChecking, and the MediaLiteracy group). For this part, we used German keywords equivalent to "protein", "vitamin", "nutrition", "diet" and "dietary supplement" to find the experimental stimuli. The main idea is to explore if the fact-checking and the media literacy interventions stay effective in a different context that is health-related, too, but unlikely to be influenced by politically motivated reasoning.[21] As before, participants in the NoIntervention, the FactChecking, and the MediaLiteracy group are shown two fakes and two facts on nutritional topics in randomized order; these fakes and facts have to fulfill the same requirements as

[15] We did not preregister whether we would display the fact-checks before or after showing the respective fake (see Appendix B for details). In our main experiment, we show the fact-check *before* the respective fake. Appendix E.3.4 shows results for a follow-up experiment where we randomize the order of fact-checks and fakes.

[16] These tips have been developed in cooperation with several professional fact-checking initiatives. See https://www.facebook.com/help/188118808357379 (viewed December 2021).

[17] See, e.g., the European Commission: https://digital-strategy.ec.europa.eu/en/policies/media-literacy (viewed June 2024).

[18] The fact-checks are self-contained and can stand on their own.

[19] More specifically, we use a quadratic scoring rule, whereby answers close to the true value increase participants' chance to receive a bonus payment of 20 EUR.

[20] We ask participants about their vaccination status at the latest possible point in the experiment to avoid priming effects. However, it is possible that the experimental stimuli could increase participants' self-reported vaccination rates. If this was true, vaccination status cannot be used as a control variable. We are not overly concerned with this possibility since in this case we would expect the vaccination rate in the PassiveControl group to be as low as in the NoIntervention group. Yet, as shown in Table A.1 in Appendix D, this is not the case. To ensure full transparency, we present the results both with and without using vaccination status as a control variable.

[21] Though not as topical as Corona vaccines, the consumption of (needless) dietary supplements is an important concern. Recent surveys indicate that nearly 50% of all German adults have purchased dietary supplements within the last six months, but almost a third of them feels ill-informed about potential health risks that go along with their consumption (Verbraucherzentrale, 2022). Moreover, the consumption of dietary supplements does typically not go along with improved public health (Radimer et al., 2004) – quite the contrary – as dietary supplements are often either ineffective (DGE, 2012) or even harmful (Chiou et al., 2011).

above. All fakes on nutrition promote the intake of needless dietary supplements such as extra protein or Vitamin C. Participants in the PassiveControl and the JustFacts group are not exposed to fakes on nutrition, but the latter receive two facts and two fact-checks.[22]

Analogous to part one of the survey, we proceed with a quiz that comprises four factual questions tailored to the fakes and facts that have just been shown:

1. How much protein should a person consume daily (in grams per kilogram bodyweight)?
2. What percentage of adults suffer from a lack of Vitamin C?
3. How many grams of microplastics does a person consume on average per week?
4. How many percent of German adolescents exercise less than recommended by the World Health Organization?

Again, each of those questions asks for a specific number, answers must be given through an input box, and we remind our participants of the potential bonus payment to incentivize high quality answers.

Finally, we inquire all participants' willingness to consume dietary supplements, where they can respond on a 5-point Likert scale ranging from *Very likely* to *Very unlikely*. To maintain participants' engagement, we placed an attention check before proceeding to the final part of the survey (see Appendix F).

Credibility. Next, we inquire the perceived credibility of all fakes, facts, and fact-checks. To this end, we display all postings again and let participants rate their credibility on a 5-point Likert scale ranging from *Very credible* to *Very incredible*. Participants are only asked about postings that they saw during the experiment, i.e., the FactChecking group is asked about fakes, facts, and fact-checks, the MediaLiteracy and the NoIntervention groups are asked about fakes and facts, the JustFacts group is asked about facts and fact-checks, and the PassiveControl group is not asked at all. We deliberately inquire the credibility of fakes, facts, and fact-checks at this late stage of the experiment to avoid priming effects on the preceding questions. Moreover, to avoid experimenter demand effects, the credibility questions are not incentivized with a potential bonus payment, and we explicitly state that there is "no correct answer" and that we are "interested in [the participants'] personal opinion".

Ex-post survey. The ex-post survey helps us better understand potential mechanisms and gather information we intentionally omitted earlier to avoid priming effects. We proceed in three steps:

(Un-)trustworthy elements. First, we investigate whether enhanced awareness and critical thinking are plausible mechanisms driving differences between treatment groups. We assess whether participants improve at considering trustworthy elements in facts (e.g., a reputable news source) and untrustworthy elements in fakes (e.g., spelling mistakes), but not vice versa, ensuring that they make conscious and correct choices. To that end, we display all fakes, facts, and fact-checks again and let participants indicate which elements of each posting they perceive as particularly trustworthy or untrustworthy.

To obtain comparable responses, we pre-define five categories of elements in the postings: (i) *format and spelling*, (ii) *information as such*, (iii) *images*, (iv) *source and URL*, and (v) *verified account*. The category *format and spelling* includes all grammar, spelling, and punctuation mistakes, as well as overly many question and exclamation marks, or uncommon fonts and colors. Elements from the category *information as such* comprise numbers and dates in a posting as well as all further "factual" information (e.g., "we are all suffering from a lack of vitamins"). *Images* refers to all photographs, graphs, and other illustrations in a posting. The category *source and URL* comprises both the author of the posting, including his or her profile name and picture, and the claimed source of the information presented, ranging from "my friend said" to "a study from MIT". Finally, the *verified account* category refers to Facebook's blue check mark. Note that some of the postings do not exhibit elements from each category; e.g., some of them feature no image, no verified account label, or no spelling mistakes. Hence, the number of elements that participants can consider as trustworthy or untrustworthy is different for each posting. See Appendix A.4 for some illustrative examples.

Our pre-defined elements are not immediately visible for the participants. Only as soon as they hover their cursor over a pre-defined element – say, a spelling mistake – the respective element gets shaded and participants can either single-click (indicate untrustworthy) or double-click (indicate trustworthy) on it. Participants can consider none, one, or several elements per posting as either trustworthy or untrustworthy. Participants must complete a mandatory tutorial to get familiar with the procedure before they can proceed. As with the credibility questions, participants are only asked about postings that they saw during the experiment (e.g., only participants in the FactChecking group are asked to consider (un-)trustworthy elements in fact-checks).

Further questions. Second, we ask further questions on participants' political party preferences, social media usage, if they got vaccinated against any type of disease during the past ten years, and if they agree with the current Corona regulations. Moreover, we ask all participants if they searched for further information online.

List experiments. Third, we conduct two list experiments à la Blair and Imai (2012) to check whether our main results on attitudes are driven by a "Bradley effect" (Hopkins, 2009), whereby participants conceal socially undesirable opinions and attitudes (see Section 6.6.3 for further details).[23]

Debriefing. At the end of Wave I, we debrief all participants by displaying the correct answers to the factual questions on Corona vaccines and nutrition.

3.1.2. Wave II

To study the effectiveness of our fact-checking and media literacy interventions in the short-run, we re-invite all participants after about one week to Wave II of the experiment. Wave II replicates the steps from Wave I, except that there are no further interventions (i.e., the setup is identical for the No Intervention, the Fact-checking, and the Media Literacy groups), no baseline survey, and that we use a different set of fakes, facts, and fact-checks.[24]

We conclude the survey with a full debriefing of all participants. To this end, we display the correct answers to all factual questions, show all fact-checks to all fakes that were used during the survey, and provide links to trustworthy websites on Corona vaccines and nutrition, where the participants can get further information on these topics if they wish.[25]

[22] Note that the FactChecking group does *not* receive these fact-checks.

[23] Our list experiments were not pre-registered.

[24] The factual questions on Covid-19 vaccines are: (i) How many European countries have completely taken the AstraZeneca Corona vaccine off the market?, (ii) In which year was the vaccine shown here manufactured?, (iii) How many doctors missed a scheduled vaccination appointment in Saarland in February?, and (iv) How many compensation claims for complications from a Corona vaccination were submitted to the state of NRW by the end of June?. The factual questions on nutrition are: (i) What percentage of Germans have a Vitamin B12 deficiency?, (ii) How much Vitamin E should an adult in Germany consume per day on average (in milligrams)?, (iii) What percentage of sugar does a German child consume on average too much?, and (iv) How many kilocalories are in 100 grams of chia seeds?

[25] Specifically, we suggest to visit the websites of the *Robert Koch Institut (RKI)* (URL: https://www.rki.de/DE/Home/homepage_node.html) and the *National Ministry of Health* (URL: https://www.bundesgesundheitsministerium.de/) further information on Corona vaccines, and to visit the website of the *German Agency for Nutrition* (URL: https://www.dge.de/) for further information on nutrition.

3.2. Implementation

The experiment was programmed with the survey software *Qualtrics* and conducted in cooperation with *respondi*, a major commercial panel provider.[26] We used e-mails to invite around 3000 German participants to Wave I of the survey, i.e., around 600 participants per group.[27] Participants had to be between 18 and 59 years old; conditional on that requirement, the sample is representative for the German population in terms of gender, age, and state of residency.[28] Participants could use their smartphones, tablets, or desktop PCs to answer our questions. Those who completed the survey received the usual payment by *respondi* plus the potential bonus payment.[29]

We conducted Wave I of the experiment between September 9th and September 29th in 2021, and Wave II between September 26th and October 27th. Participants received a re-invitation about one week after they completed Wave I. The minimum interval of actual participation in the two waves is equal to eight days, though, the median interval is equal to 15, and the mean interval equal to 17.6 days (see Section 6.6.2 for further discussion). The response rate in Wave II is equal to 83% – which is roughly equal to *respondi*'s average – and we find no evidence for differential attrition.[30] At the time the experiment took place, all German adults had the opportunity to get fully vaccinated (two injections), and policy makers and health experts were discussing whether and when a third injection would make sense.

3.3. Balance check

Table A.1 in Appendix D displays the means and standard deviations of all control variables for each treatment group. Since we use the NoIntervention group as baseline in the subsequent analyses, we also conduct t-tests on the difference in means between the NoIntervention and each of the other treatment groups, respectively.

We find that our sample is strongly balanced with respect to age, gender, family status, state of residence, consumption of dietary supplements, and prior knowledge on current events, health, and nutrition, but there are small differences between some of the treatment groups for household income, education, party preferences, and Corona vaccination status. To take these imbalances into account, we include the full set of pre-registered control variables into each of our regression analyses.[31] Since we did not pre-register participants' Corona vaccination status as a control, we only include it as a robustness check – with one exception (see Section 4.1.3) our results are unaffected.[32]

3.4. Variables

Next, we aggregate our participants' responses to the various fakes and facts and convert them into measures suitable for regression analyses. We also standardize responses to the prior knowledge questions and generate an indicator for participants' uncertainty about them. Table A.2 provides summary statistics of all dependent variables that we use in the analysis.[33]

Credibility. We start by computing each participant's mean response to the credibility questions on Corona vaccine fakes, Corona vaccine facts, nutrition fakes, and nutrition facts for each of the two waves, respectively (i.e., we compute eight mean responses per participant). Then, we define a dummy variable equal to one if the mean response indicates that the participant perceives the fakes or facts on average as *Very credible*, *Credible* or *Indecisive*.[34] This aggregation level allows us to examine the treatment effect on fakes and facts separate from each other, whereby we can show that our interventions have no detrimental effect on facts.

Factual knowledge. Next, we standardize participants' responses to the factual knowledge questions. To this end, we first compute the absolute distance between each response and the correct answer. E.g., the correct answer to "How many people died after vaccination in a Sana clinic?" is equal to zero; if the participant's response is "50", the absolute distance between response and correct answer is equal to 50. To avoid distortion through outliers, we winsorize all distances to each question at their 95th percentile.[35] Then, based on the entire sample, we standardize all winsorized distances to have a mean of zero and a standard deviation of one, which allows us to compare responses across questions. Finally, we aggregate participants' responses by computing their mean standardized distance to the correct answer to the factual questions on Corona vaccine fakes, Corona vaccine facts, nutrition fakes, and nutrition facts for each of the two waves, respectively (i.e., we compute eight mean responses per participant again).

Attitudes. To capture participants' attitudes, we define a dummy that is equal to one if participant i states to be *Very likely*, *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19 in each of the two waves, respectively. Analogously, we define a dummy equal to one if he or she states to be *Very likely*, *Likely* or *Indecisive* to consume dietary supplements in the near future.[36]

(Un-)trustworthy elements in fakes and facts. To measure how much attention participants pay to the content of the fakes and facts and how critically they evaluate them, we count how many elements (in absolute terms) they consider as trustworthy or untrustworthy in each posting. Then, for each participant, we compute the average number of (un-)trustworthy elements that he or she considered for Corona vaccine fakes, Corona vaccine facts, nutrition fakes, and nutrition facts for each of the two waves, respectively.

[26] Cooperating with professional panel providers such as *respondi* has become standard in economic research; see, e.g., Stantcheva (2021) and Alesina et al. (2023) for examples and https://www.bilendi.com/ for further details on *respondi*.

[27] The sample size is comparable to related studies, e.g. Barrera et al. (2020) and Henry et al. (2022).

[28] Although our participants are likely to encounter misinformation on Corona vaccines frequently in their every day lives, we exclude participants aged 60+ as an especially vulnerable group from our experiment.

[29] Our sample size and expected attrition rate were preregistered; see Appendix B for details.

[30] Response rates per group: NoIntervention 82.36%, FactChecking 84.51%, MediaLiteracy 83.36%, JustFacts 78.90%, PassiveControl 84.62%.

[31] The preregistered control variables are age, gender, family status, state of residence, personality traits ("big 5"), household income, education, party preferences, and prior knowledge on current events, health, and nutrition; see also Appendix B.

[32] For a detailed discussion of heterogeneity in terms of vaccination status see Section 6.6.1.

[33] We preregistered our main outcome variables as such, but not that we would measure them on a 5-point Likert scale and transform them to binary outcomes or standardized values (see Appendix B for details). However, our results are robust to alternative specifications of the outcome variables.

[34] Our results are robust to alternative cutoffs and to using the average response on the 5-point Likert scale prior to its transformation to a binary measure. Even with an extreme cutoff, where the indicator variable is set to 1 only if participants rated fakes as *very incredible*, the findings remain consistent (taking into account the reverse coding).

[35] Note that we pre-registered our intention to winsorize each participant's responses at their 95th percentile. We obtain similar results when we drop outliers beyond the 95th percentile, winsorize responses at their 99th percentile, or do not winsorize at all. The distribution is based on the entire sample.

[36] Our results are robust to alternative cutoffs and to using the average response on the 5-point Likert scale prior to its transformation to a binary measure.

Prior knowledge. Analogous to factual knowledge, we make participants' responses to each of the three prior knowledge questions better comparable by computing the standardized distance between a participant's response and the correct answer. In addition, we compute an indicator that is equal to one if participant i is on average *Very uncertain*, *Uncertain*, or *Somewhat certain* about his or her prior knowledge.

3.5. ITT analysis

In our baseline analysis, we estimate the Intention to Treat Effects (ITT) of our interventions by using OLS to estimate the regression equation

$$y_{iw} = \beta_0 + \beta_1 TG_i + \beta_2 X_i + \varepsilon_{iw}, \quad (1)$$

where y_{iw} corresponds to an outcome of participant i in survey wave w as described above, TG_i denotes participant i's treatment group, and X_i is a vector of pre-registered control variables including age, gender, party preferences, religion, education, family status, household income, personality traits, state of residence, and prior knowledge on current events, health, and nutrition. The baseline category in TG_i is the NoIntervention group, i.e., we compare participants who receive fakes and facts without further intervention to participants in each of the other treatment groups.[37]

We refer to these estimates as ITT, because participants from the FactChecking and the MediaLiteracy group could theoretically skip the intervention by just quickly clicking through the survey. We show in Section 6.6.4 that our findings are robust to an IV-specification.

4. Results

The main purpose of our paper is to study whether and when fact-checking and media literacy interventions are able to debunk fake news that circulate on social media. To this end, we focus on comparing the NoIntervention to the FactChecking and the MediaLiteracy group here and defer supporting analyses of the JustFacts and the PassiveControl group to Section 6.

We consider three types of outcomes: the credibility of fakes and facts, factual knowledge on the topics the fakes and facts are dealing with, and attitudes towards Corona vaccination and the intake of dietary supplements. The idea is to examine a coherent cognitive chain: Do the interventions reduce the credibility of fakes (but not of facts)? If yes, does that translate into better factual knowledge on the topics the fakes are dealing with? If yes, does that affect participants' attitudes? Table 1 displays all mean outcomes per treatment group.

4.1. ITT analysis

We start by examining the ITT of our interventions (Section 6.6.4 presents an IV analysis). In particular, we demonstrate that the effectiveness of fact-checking tends to be limited to the fakes that are corrected, while the media literacy intervention helps to distinguish between fakes and facts more generally. Figures A.8 to A.10 in Appendix C illustrate the results for each outcome, treatment, and wave of the survey; further details are presented below.

[37] We preregistered our baseline OLS regression estimation as described above. See Appendix B for details.

4.1.1. Credibility

Table 2 presents the regression results for the perceived credibility of fakes. Panel A shows the estimates from comparing the FactChecking, and Panel B from comparing the MediaLiteracy to the NoIntervention group, respectively.[38]

Our first main result is that the fact-checking intervention reduces the credibility of fakes on Corona vaccines in Wave I of the survey. In Wave I, participants from the FactChecking group are on average 7 to 8 percentage points less likely to perceive fakes on Corona vaccines on average as credible than participants from the NoIntervention group. The estimate is statistically significant at the 1%-level, the effect size corresponds to about 15.5% of a standard deviation in the dependent variable in the baseline NoIntervention group and to about 23.7% of its mean value. In contrast to that, we find no statistically significant differences between the FactChecking and the NoIntervention group for fakes that are not corrected by fact-checks, i.e., fakes on Corona vaccines in Wave II of the survey and fakes on nutrition in either wave.

Second, we find that the media literacy intervention reduces the credibility of fakes more generally than fact-checking. In Wave I, participants from the MediaLiteracy group are about 10 percentage points less likely to consider fakes on Corona vaccines on average as credible than participants from the NoIntervention group. The estimate is statistically significant at the 1%-level; the effect size corresponds to about 22.9% of a standard deviation in the dependent variable and to 35% of its baseline value, whereby the effect is even larger than for the FactChecking group.[39] Unlike fact-checking, the media literacy intervention also reduces the credibility of all fakes on nutrition and of fakes on Corona vaccines in Wave II of the survey, although the latter effect is small and not statistically significant when we include our controls. The estimates for nutrition, however, correspond to about 18.1% of a standard deviation in the dependent variable in the baseline NoIntervention group and to about 8.2% of its mean value in both waves of the survey.[40]

Crucially, neither intervention reduces the credibility of facts (see Table A.3 in Appendix D), i.e., participants do not become more skeptical towards social media postings per se. Instead, our results indicate that the fact-checking and, in particular, the media literacy intervention enhance participants' truth discernment (Pennycook and Rand, 2021), whereby they can better distinguish between false and correct information that they encounter online.[41]

[38] Tables A.7 to A.9 in Appendix D show the results for a direct comparison of the FactChecking and the MediaLiteracy group.

[39] The difference between the FactChecking and the MediaLiteracy group is not statistically significant, though (two-sided t-test, $p = 0.176$).

[40] Note that the mean perceived credibility of fakes on Corona vaccines is much lower than for fakes on nutrition. To analyze whether fact-checking more credible fakes enhances participants' critical evaluation of subsequent posts, we varied the exposure by presenting fakes with a higher mean baseline credibility in a follow-up study, which were fact-checked before participants proceeded to a new topic without further intervention. For a detailed discussion of the findings see Appendix E.3.

[41] Participants in the MediaLiteracy group might expect to see a certain proportion of fakes, leading to potential experimenter demand effects. However, participants did not know the total number of news items, preventing specific inferences about the total number of fakes they would encounter. Additionally, the results in Tables A.4 and A.6 in Appendix D indicate that participants in the MediaLiteracy group improved in distinguishing false from correct news and in considering trustworthy elements in facts and untrustworthy elements in fakes. Further, Table A.30 in Appendix E.3.6 shows that particpants in the MediaLiteracy group in the follow-up experiment also spend more time reading the posts. This supports the idea that their judgments are based on enhanced skills rather than an expectation of a certain proportion of fakes. Finally, our list experiments in Section 6.6.3 show no evidence of greater experimenter demand effects in the MediaLiteracy group compared to other groups.

Table 1
Mean values of all outcomes per treatment group

Panel A: Wave I	**Credibility**		**Knowledge**		**Truth discernment** Credibility		Knowledge		**Attitudes**	
	Covid	Nutrition	Covid	Nutrition	Covid	Nutrition	Covid	Nutrition	Covid	Nutrition
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
No Intervention	0.299	0.901	0.398	0.132	0.584	0.922	0.152	0.016	0.783	0.557
	(0.018)	(0.012)	(0.034)	(0.025)	(0.014)	(0.008)	(0.024)	(0.020)	(0.017)	(0.020)
Fact-checking	0.219	0.886	0.076	0.051	0.538	0.913	0.021	−0.032	0.830	0.555
	(0.017)	(0.013)	(0.033)	(0.025)	(0.014)	(0.008)	(0.023)	(0.019)	(0.015)	(0.020)
Media Literacy	0.195	0.842	0.179	0.053	0.529	0.896	0.047	−0.013	0.836	0.515
	(0.016)	(0.015)	(0.034)	(0.026)	(0.014)	(0.009)	(0.023)	(0.020)	(0.015)	(0.020)
Panel B: Wave II	**Credibility**		**Knowledge**		**Truth discernment** Credibility		Knowledge		**Attitudes**	
	Covid	Nutrition	Covid	Nutrition	Covid	Nutrition	Covid	Nutrition	Covid	Nutrition
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
No Intervention	0.403	0.831	0.334	0.143	0.635	0.896	0.077	−0.031	0.741	0.554
	(0.022)	(0.017)	(0.037)	(0.026)	(0.015)	(0.010)	(0.025)	(0.021)	(0.019)	(0.022)
Fact-checking	0.394	0.806	0.286	0.164	0.625	0.872	0.081	−0.022	0.797	0.565
	(0.022)	(0.018)	(0.039)	(0.028)	(0.015)	(0.010)	(0.025)	(0.022)	(0.018)	(0.022)
Media Literacy	0.350	0.759	0.220	0.117	0.615	0.861	0.036	−0.047	0.804	0.532
	(0.021)	(0.019)	(0.037)	(0.028)	(0.015)	(0.011)	(0.025)	(0.022)	(0.018)	(0.022)

Notes: Table 1 shows the mean outcome values for each of our treatment groups, respectively. Standard errors in parentheses.

Table 2
Credibility of fakes.

Panel A: Fact-checking	**Wave I**				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fact-checking	−0.080	−0.071	−0.015	−0.010	−0.009	0.016	−0.026	−0.018
	[0.025]	[0.024]	[0.018]	[0.017]	[0.031]	[0.030]	[0.024]	[0.024]
p-value	(0.001)	(0.004)	(0.385)	(0.553)	(0.769)	(0.582)	(0.282)	(0.460)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,225	1,225	1,225	1,225	1,022	1,022	1,022	1,022
Panel B: Media literacy	**Wave I**				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Media literacy	−0.104	−0.105	−0.060	−0.061	−0.052	−0.042	−0.072	−0.068
	[0.024]	[0.024]	[0.019]	[0.019]	[0.030]	[0.029]	[0.025]	[0.025]
p-value	(0.000)	(0.000)	(0.002)	(0.001)	(0.084)	(0.143)	(0.004)	(0.006)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,231	1,231	1,231	1,231	1,020	1,020	1,020	1,020
Baseline: No Intervention								
Mean DV	0.299	0.299	0.901	0.901	0.403	0.403	0.831	0.831
Std.Dev. DV	0.458	0.458	0.299	0.299	0.491	0.491	0.375	0.375

Notes: Table 2 shows the OLS estimates of comparing the NoIntervention to the FactChecking (Panel A) and to the MediaLiteracy group (Panel B), respectively. The dependent variable is a dummy equal to one if participant i perceives the **fakes** on Corona vaccines and nutrition in Wave I and in Wave II of the survey on average as *Very credible*, *Credible* or *Indecisive*. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition.

Truth discernment. This result is further confirmed by an explicit analysis of truth discernment. In particular, we consider fakes and facts on Corona vaccines (nutrition) in Wave I (Wave II) of the survey in a single regression equation and estimate

$$cred_{ij} = \alpha_0 + \alpha_1 fake_j + \alpha_2 TG_i + \alpha_3 (TG_i * fake_j) + \theta X_i + e_{ij} \quad (2)$$

by OLS, where $cred_{ij}$ is a dummy equal to one if participant i perceived group j of news items (fakes or facts) on average as *Very credible*, *Credible* or *Indecisive*, TG_i is a treatment indicator with the NoIntervention group as omitted category, and $fake_j$ indicates fakes (in contrast to facts). The parameter α_1 thus corresponds to the average difference in the perceived credibility of fakes and facts – which we interpret as truth discernment – in the baseline NoIntervention group. The parameter of interest is α_3, which corresponds to the average difference in truth discernment between the NoIntervention and the FactChecking or the MediaLiteracy group, respectively.

Table A.4 in Appendix D displays our results. Panel A shows that relative to facts, participants from the FactChecking group are on average 6.8 percentage points less likely to perceive fakes on Corona vaccines as credible than participants from the NoIntervention group. Hence, fact-checking enhances participants' truth discernment for the fakes that are explicitly targeted by our intervention. In contrast to that, we do not observe a statistically significant difference in truth discernment between the FactChecking and the NoIntervention group for Corona vaccines in Wave II of the survey or for nutrition in either wave.

Table 3
Distance to truth on topics covered by fakes.

	Wave I				Wave II			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Fact-checking								
Fact-checking	−0.322	−0.319	−0.080	−0.080	−0.048	−0.051	0.021	0.030
	[0.047]	[0.048]	[0.035]	[0.035]	[0.054]	[0.054]	[0.038]	[0.038]
p-value	(0.000)	(0.000)	(0.023)	(0.024)	(0.371)	(0.350)	(0.596)	(0.426)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,225	1,225	1,225	1,225	1,022	1,022	1,022	1,022
Panel B: Media literacy								
	Wave I				Wave II			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Media literacy	−0.219	−0.220	−0.078	−0.081	−0.114	−0.138	−0.026	−0.041
	[0.048]	[0.048]	[0.037]	[0.036]	[0.053]	[0.052]	[0.038]	[0.037]
p-value	(0.000)	(0.000)	(0.032)	(0.027)	(0.031)	(0.009)	(0.488)	(0.264)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,231	1,231	1,231	1,231	1,020	1,020	1,020	1,020
Baseline: No Intervention								
Mean DV	0.299	0.299	0.901	0.901	0.403	0.403	0.831	0.831
Std.Dev. DV	0.458	0.458	0.299	0.299	0.491	0.491	0.375	0.375

Notes: Table 3 compares distance to truth on topics that the Corona vaccine and nutrition **fakes** are dealing with between participants from the FactChecking (Panel A) and the MediaLiteracy (Panel B) and the NoIntervention group, respectively. All estimates are OLS estimates. The dependent variable is equal to participant *i*'s average standardized distance to the correct answer. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition.

Unlike fact-checking, the media literacy intervention enhances participants' truth discernment throughout the entire survey (Panel B). In particular, relative to facts, participants from the MediaLiteracy group are on average between 6.6 and 9.8 percentage points less likely to perceive fakes on Corona vaccines on average as credible than participants from the NoIntervention group. With the exception of Corona vaccines in Wave II of the survey, all estimates are highly statistically significant at the 1%-level.

4.1.2. Factual knowledge

Table 3 shows the regression results for participants' factual knowledge on the topics the fakes are dealing with. Again, Panel A shows the estimates from comparing the FactChecking, and Panel B from comparing the MediaLiteracy to the NoIntervention group, respectively.

Consistent with the results on credibility, Panel A shows that the fact-checking intervention enhances participants' factual knowledge on Corona vaccines in Wave I of the survey. Specifically, responses by the FactChecking group are on average about 0.32 standard deviations closer to the correct answer than responses by the NoIntervention group; the effect is statistically significant at the 1%-level.[42] Somewhat surprisingly, we also observe that participants from the FactChecking group give better answers to the factual knowledge questions on nutrition in Wave I of the survey. According to our estimates, responses by the FactChecking group are about 0.08 standard deviations closer to the correct answer than responses by the NoIntervention group; the effect is statistically significant at the 5%-level. A potential explanation is that the presence of fact-checking makes participants generally more cautious towards implausible information, although it does not affect the perceived credibility of fakes that are not explicitly corrected. This would be in line with recent findings by Barrera et al. (2020) and Nyhan et al. (2020), who show that fact-checks can improve the accuracy of respondents' factual beliefs, but fail to affect more deep-rooted perceptions and attitudes (see Section 4.1.3 for further discussion). Consistent with that, we find no statistically significant differences in factual knowledge on fakes between the FactChecking and the NoIntervention group in Wave II of the survey, where no further fact-checks are shown.

Analogous to the results on credibility, Panel B shows that the media literacy intervention enhances participants' factual knowledge more generally than fact-checking. In Wave I of the survey, responses by the MediaLiteracy group are about 0.22 standard deviations closer to the correct answer than responses by the NoIntervention group for fakes on Corona vaccines, and about 0.08 standard deviations closer for fakes on nutrition. Both effects are statistically significant. The impact of the media literacy is thus smaller than the impact of the fact-checking intervention when the FactChecking group receives correct information in addition to the fakes, and roughly equivalent when it does not.[43] However, unlike fact-checking, the media literacy intervention could also improve participants' factual knowledge on Corona vaccines in Wave II of the survey. Specifically, responses by the MediaLiteracy group are about 0.14 standard deviations closer to the correct answer than responses by the NoIntervention group; the effect is statistically significant at the 1%-level when we include our controls. In contrast to that, we find no statistically significant difference in factual knowledge on nutrition between the MediaLiteracy and the NoIntervention group in Wave II of the survey, although the estimates have the expected sign. One plausible explanation is that, according to Table 1, fakes on nutrition seem to be more credible on average than fakes on Corona vaccines. As a result, the tips to spot false news could be more difficult to apply, which in turn entails a smaller difference between the MediaLiteracy and the NoIntervention group. In addition, the impact of our intervention is likely to decay over time (e.g., Nyhan, 2021; Maertens et al., 2021), which further reduces the effect size in Wave II of the survey.

Similar to the results on credibility, Table A.5 in Appendix D shows that neither intervention reduces participants' factual knowledge on

[42] Note that aggregating the standardized responses to Corona vaccine fakes, Corona vaccine facts, nutrition fakes, and nutrition facts in each wave of the survey causes the reported means and standard deviations in Table 3 to be unequal to zero and one, respectively.

[43] The difference between the FactChecking and the MediaLiteracy group is statistically significant at the 5%-level for fakes on Corona vaccines in Wave I of the survey (two-sided *t*-test, $p = 0.034$) and weakly statistically significant at the 10%-level for fakes on nutrition in Wave II of the survey (two-sided *t*-test, $p = 0.097$).

Table 4
Attitudes towards Corona vaccination and the intake of dietary supplements.

Panel A: Fact-checking										
	Wave I					**Wave II**				
	Corona vaccination			Supplements		Corona vaccination			Supplements	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Fact-checking	0.047	0.038	0.017	−0.001	0.003	0.057	0.048	0.034	0.011	0.021
	[0.023]	[0.021]	[0.019]	[0.028]	[0.028]	[0.026]	[0.025]	[0.022]	[0.031]	[0.031]
p-value	(0.037)	(0.068)	(0.341)	(0.959)	(0.908)	(0.032)	(0.049)	(0.127)	(0.717)	(0.504)
Controls	no	yes	yes +	no	yes	no	yes	yes +	no	yes
N	1,225	1,225	1,225	1,225	1,225	1,022	1,022	1,022	1,022	1,022
Panel B: Media literacy										
	Wave I					**Wave II**				
	Corona vaccination			Supplements		Corona vaccination			Supplements	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Media literacy	0.054	0.053	0.034	−0.041	−0.037	0.064	0.068	0.048	−0.022	−0.012
	[0.022]	[0.021]	[0.019]	[0.028]	[0.028]	[0.026]	[0.025]	[0.022]	[0.031]	[0.031]
p-value	(0.016)	(0.012)	(0.077)	(0.148)	(0.188)	(0.015)	(0.006)	(0.033)	(0.486)	(0.696)
Controls	no	yes	yes +	no	yes	no	yes	yes +	no	yes
N	1,231	1,231	1,231	1,231	1,231	1,020	1,020	1,020	1,020	1,020
Baseline: No Intervention										
Mean DV	0.813	0.813	0.813	0.548	0.548	0.781	0.781	0.781	0.552	0.552
Std.Dev. DV	0.390	0.390	0.390	0.500	0.500	0.413	0.413	0.413	0.497	0.497

Notes: Table 4 presents the OLS estimates of comparing the NoIntervention to the FactChecking (Panel A) and to the MediaLiteracy group (Panel B), respectively. The dependent variable is a dummy equal to one if participant *i* states to be *Very likely*, *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19, or *Very likely*, *Likely* or *Indecisive* to consume dietary supplements in the near future. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition. In columns 3 and 8 ("yes +"), we also control for participants' Corona vaccination status.

topics that the facts are dealing with. Hence, both the fact-checking and the media literacy intervention enhance participants' factual knowledge on average. Moreover, Table A.6 shows that the fact-checking intervention enhances participants' truth discernment for Corona vaccines in Wave I of the survey, while there is no statistically significant difference in truth discernment between the FactChecking and the NoIntervention group otherwise. The media literacy intervention, in contrast, enhances truth discernment more generally. In particular, all estimates have the expected sign and almost all of them are statistically significant.

4.1.3. Attitudes

Table 4 displays the regression results for participants' attitudes towards Corona vaccination and the intake of (needless) dietary supplements. Panel A displays the estimates from comparing the FactChecking, and Panel B from comparing the MediaLiteracy to the NoIntervention group, respectively.

Panel A reveals that the impact of fact-checking on participants' attitudes is extremely limited. Although participants from the FactChecking group are more likely to state that they are *Very likely*, *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19 than participants from the NoIntervention group, much of the effect is driven by the smaller proportion of fully vaccinated participants in the latter (see Section 3.3). In particular, we find that a participant's decision to get vaccinated in the past strongly predicts his or her intention to get vaccinated in the future. As a result, the estimated difference in the average willingness to get vaccinated between the FactChecking and the NoIntervention group shrinks and becomes statistically insignificant when we control for participants' Corona vaccination status in columns 3 and 8.[44] Similarly, we do not find any statistically significant difference in participants' willingness to consume dietary supplements in either wave of the survey. We must therefore conclude that the fact-checking intervention – though effective in reducing the perceived credibility of and enhancing factual knowledge on the fakes that are being targeted – fails to affect participants' attitudes on average. This is consistent with earlier findings by Barrera et al. (2020), Swire et al. (2017), Nyhan et al. (2020), and Jerit and Zhao (2020), among others, who show that fact-checking can help to create "a more informed citizenry" (Nyhan et al., 2020, p.942), but struggles to change more deep-rooted perceptions and attitudes such as which political party to support or, as in our context, whether to get vaccinated against Covid-19 or not.

In line with the results on credibility and factual knowledge, Panel B shows that the media literacy intervention is effective in swaying participants' attitudes. In particular, participants from the MediaLiteracy group are 3.4 to 4.8 percentage points more likely to state that they are *Very likely*, *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19 than participants from the NoIntervention group. In contrast to fact-checking, this difference remains statistically significant when we control for participants' Corona vaccination status.[45] The effect size corresponds to about 8.7% of a standard deviation in the dependent variable in the baseline NoIntervention group and 4.2% of its mean value in Wave I of the survey, and to about 11.6% of a standard deviation in the dependent variable and 6.1% of its baseline value in Wave II. Roughly 85% of all participants report that they are willing to get vaccinated or boostered against Covid-19, though. The relatively small effect size could thus be explained by a small proportion of participants who can still be convinced to get the shot; Section 4.3 computes persuasion rates à la DellaVigna and Kaplan (2007) to further address this issue. The estimates for participants' willingness to consume needless dietary supplements are statistically insignificant. One potential explanation is that, unlike Corona vaccination, the intake of dietary supplements is typically based on year-long habits (Bailey et al., 2013), and even if the media literacy intervention could affect participants' attitudes, further effects from attitudes to habit change are typically modest (Verplanken and Orbell, 2022).

In sum, the results from Section 4.1 support the idea that the effect of fact-checking tends to be limited to the fakes that are being corrected, while enhancing participants' media literacy helps them

[44] Note that the willingness to get vaccinated or boostered against Covid-19 is the only instance where the smaller proportion of fully vaccinated participants in the NoIntervention group plays a role.

[45] The differences between the FactChecking and the MediaLiteracy group are not statistically significant, though.

Table 5
Romano-Wolf Multiple Hypothesis Correction.

Panel A: Fact-checking	**Wave I Corona**			**Nutrition**			**Wave II Corona**			**Nutrition**		
	Cred.	Dist.	Vaccine	Cred.	Dist.	Suppl.	Cred.	Dist.	Vaccine	Cred.	Dist.	Suppl.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original *p*-value	0.004	0.000	0.341	0.553	0.024	0.909	0.582	0.350	0.127	0.460	0.426	0.504
Romano-Wolf *p*-value	0.070	0.000	0.973	0.973	0.304	0.973	0.973	0.973	0.802	0.973	0.973	0.973
Panel B: Media literacy	**Wave I Corona**			**Nutrition**			**Wave II Corona**			**Nutrition**		
	Cred.	Dist.	Vaccine	Cred.	Dist.	Suppl.	Cred.	Dist.	Vaccine	Cred.	Dist.	Suppl.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original *p*-value	0.000	0.000	0.077	0.001	0.027	0.188	0.143	0.009	0.033	0.006	0.264	0.700
Romano-Wolf *p*-value	0.001	0.000	0.642	0.028	0.323	0.878	0.816	0.134	0.354	0.097	0.948	0.973

Notes: Table 5 displays the original and the Romano-Wolf p-values from our ITT analysis. The Romano-Wolf p-values were obtained with the STATA package *rwolf2* (Clarke et al., 2020). Each specification includes the full set of control variables. In columns 3 and 9, we also include participants' vaccination status as additional control.

to distinguish between fakes and facts more generally. Hence, in an environment where not every message can be fact-checked, media literacy interventions are likely to be more effective on average.

4.1.4. Multiple hypothesis testing

Our ITT analysis examines two treatments and twelve different outcome variables; i.e., we consider twenty-four hypotheses in sum. To take potential over-rejection of null hypotheses into account, we conduct a Romano-Wolf multiple hypothesis correction (Romano and Wolf, 2005). This procedure uses resampling methods (e.g., bootstrap) to control for the so-called familywise error rate, i.e., the probability of rejecting at least one true null hypothesis in the family of hypotheses under test. The Romano-Wolf procedure offers more power than the procedures by Bonferroni and Holm, and it is furthermore able to eliminate the so-called subset pivotality assumption from previous resampling-procedures such as Westfall-Young (Clarke et al., 2020).

Table 5 displays both the original and the Romano-Wolf p-values from our ITT analysis of the fact-checking (Panel A) and the media literacy intervention (Panel B), respectively. As expected, all p-values grow under the correction. Yet, our main results remain qualitatively intact in that the fact-checking intervention is limited to the fake news that are targeted, while the media literacy intervention helps more generally. In particular, we observe that the media literacy intervention has a statistically significant impact on the credibility of both fakes on Covid-19 vaccines and nutrition in Wave I of the survey even after the Romano-Wolf correction, and that one of the effects in Wave II survives.

4.2. Heterogeneity in baseline beliefs

The average impact of our fact-checking and media literacy interventions is likely to depend on participants' prior beliefs. In particular, if participants update their beliefs in a Bayesian fashion, those with prior beliefs that are most different from the presented information would change their beliefs the most (see, e.g., Fong et al., 2024, for empirical evidence). This section demonstrates that our interventions are indeed more effective for supporters of the AfD ("Alternative for Germany"), a far-right populist party known for spreading misinformation on Corona vaccines (e.g., Gensing, 2021).[46] Specifically, we show that there is substantial effect heterogeneity between AfD and non-AfD supporters for fakes on Corona vaccines, but not for fakes on nutrition, where participants' prior beliefs are much more alike. Given that the effect heterogeneity is limited to fakes on Corona vaccines, we focus on that topic here, and defer the results on nutrition to Tables A.10 and A.11 in Appendix D.[47]

4.2.1. Fact-checking

Table 6 displays the average impact of fact-checking for AfD (Panel A) and non-AfD supporters (Panel B) on each of our main outcomes in each wave of the survey, respectively.

There are two main insights. First, each point estimate in Panel A is larger than its counterpart in Panel B, which means that the average impact of fact-checking is stronger for AfD than for non-AfD supporters. With the exception of factual knowledge, each of these differences is highly statistically significant (two-sided t-tests, $p < 0.001$). Second, there is ample heterogeneity in the baselines, i.e., the mean dependent variables in the NoIntervention group, which we interpret as participants' average prior beliefs: AfD supporters are on average almost twice as likely to perceive fakes on Corona vaccines as *Very credible*, *Credible* or *Indecisive* than non-AfD supporters, their responses to the factual knowledge questions are further away from the correct answer by roughly a third, and they are only half as likely to state that they are willing to get vaccinated or boostered against Covid-19. Hence, one explanation for the effect heterogeneity between AfD and non-AfD supporters is that the proportion of participants who can still update their beliefs is substantially larger for the former than for the latter group. We elaborate on this idea in Section 4.3, where we compute

[46] E.g., The AfD politician Björn Höcke claimed that Corona vaccines could cause infertility (https://www.youtube.com/watch?v=X1Gw0k9CxUY&t=1346s, viewed April 2024), Joachim Kuhs, also AfD politician, asserted that more people died from Corona vaccination in 2021 than people died from any kind of vaccination in the entire past twenty years (https://www.youtube.com/watch?v=JmbLGTH4PDw, viewed April 2024), and the AfD Bavaria denies that vaccines protect against Covid-19 (https://www.tagesschau.de/faktenfinder/afd-angst-impfungen-101.html, viewed April 2024).

[47] We do not find any systematic effect heterogeneity in terms of education, age, social media usage, support of policy measures to counteract the spread of the Corona virus, or prior knowledge on current events, health, and nutrition. When we compare participants who are fully vaccinated to those who are not, we find a similar, though less pronounced, pattern as for AfD supporters (see Section 6.6 for a detailed analysis). We preregistered a subgroup analysis that would exploit differences in participants' prior beliefs. While we initially thought that vaccination status or consent with the existing Corona remedies would be a good proxy, it turned out later that AfD support is seemingly even closer related to differences in prior beliefs. Hence, we preregistered the former but not the latter subgroup analysis. We also preregistered a subgroup analysis w.r.t. to time spent on social media, but, as argued, found no hetereogeneous effects. See Appendix B for further details.

Table 6
Heterogeneity in baseline beliefs on Corona vaccination – FactChecking.

	Wave I			Wave II		
	Cred.	Dist.	Vaccine	Cred.	Dist.	Vaccine
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Fact-checking – AfD supporters						
Fact-checking	−0.166	−0.362	0.137	−0.221	−0.325	0.090
	[0.099]	[0.171]	[0.062]	[0.096]	[0.178]	[0.081]
p-value	(0.095)	(0.036)	(0.029)	(0.024)	(0.072)	(0.267)
Controls	yes	yes	yes +	yes	yes	yes +
N	114	115	115	99	99	99
Baseline: No Intervention						
Mean DV	0.508	0.509	0.390	0.788	0.358	0.404
Std.Dev. DV	0.504	0.847	0.492	0.412	0.795	0.495
Panel B: Fact-checking – non-AfD supporters						
	Wave I			Wave II		
	Cred.	Dist.	Vaccine	Cred.	Dist.	Vaccine
	(1)	(2)	(3)	(4)	(5)	(6)
Fact-checking	−0.055	−0.316	0.006	0.046	−0.032	0.030
	[0.025]	[0.051]	[0.019]	[0.031]	[0.058]	[0.023]
p-value	(0.029)	(0.000)	(0.752)	(0.145)	(0.576)	(0.192)
Controls	yes	yes	yes +	yes	yes	yes +
N	1,107	1,110	1,110	923	923	923
Baseline: No Intervention						
Mean DV	0.277	0.386	0.825	0.369	0.331	0.779
Std.Dev. DV	0.448	0.840	0.381	0.483	0.850	0.415

Notes: Table 6 displays the effect heterogeneity between AfD supporters (Panel A) and non-AfD supporters (Panel B) for our Fact-checking intervention. In columns 1 and 4, the dependent variable is a dummy equal to one if participant *i* perceives the fakes on Corona vaccines as *Very credible*, *Credible* or *Indecisive* on average. In columns 2 and 5, the dependent variable is equal to participant *i*'s average standardized distance to the correct answer. In columns 3 and 6, the dependent variable is a dummy equal to one if participant *i* states to be *Very likely*, *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19. All estimates are OLS estimates. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition. In columns 3 and 6 ("yes +"), we also control for participants' Corona vaccination status.

persuasion rates for each of our interventions. Differing trust in the fact-checks, in contrast, does not seem to drive the effect heterogeneity: AfD supporters perceive the fact-checks on Corona vaccines on average as slightly less credible as non-AfD supporters (see Section 6.4).

Column 3 shows that AfD supporters from the FactChecking group are 13.7 percentage points more likely to state that they are willing to get vaccinated or boostered than AfD supporters from the NoIntervention group. The effect is statistically significant at the 5%-level; it corresponds to 27.8% of a standard deviation in the dependent variable in the baseline NoIntervention group and to 35.1% of its mean value. This result is especially remarkable given that fact-checking typically fails to affect participants' attitudes (see Section 4.1.3). Here, the relatively strong impact of fact-checking on the credibility of and factual knowledge on fakes, combined with the initially small average proportion of AfD supporters who want to get vaccinated or boostered (39.0% vs. 82.5% for the non-AfD supporters), is potent enough to sway attitudes of AfD supporters, while attitudes of non-AfD supporters remain unaffected.

4.2.2. Media literacy

Analogous to Table 6, Table 7 displays the average impact of the media literacy intervention for AfD supporters (Panel A) and non-AfD supporters (Panel B) on each of our main outcomes in each wave of the survey, respectively. Again, the impact of our intervention is larger for AfD supporters than for non-AfD supporters; with the exception of factual knowledge, each of these differences is highly statistically significant (two-sided *t*-tests, $p < 0.001$). We also observe large differences in participants' baselines. Hence, part of the effect heterogeneity can be explained by the different proportion of participants who can still update their beliefs (see Section 4.3 for further discussion).

Similar to fact-checking, we find that the media literacy intervention has a large positive impact on AfD supporters' attitudes towards Corona vaccination. Specifically, AfD supporters from the MediaLiteracy group are about 14.9 percentage points more likely to state that they are *Very likely* or *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19 than AfD supporters from the NoIntervention group. The effect is statistically significant at the 5%-level; it corresponds to about 30.3% of a standard deviation in the dependent variable in the baseline NoIntervention group and to 38.2% of its mean value. In contrast to fact-checking, this difference remains statistically significant in Wave II of the survey, and the effect size is comparable to Wave I. This suggests that the media literacy intervention is even more successful in swaying participants' attitudes. Though, none of the differences in means between the FactChecking and the MediaLiteracy group are statistically significant when evaluated using a two-sided t-test.

In sum, we find that both the fact-checking and the media literacy intervention are more effective for AfD supporters, whose prior beliefs on Corona vaccines are on average further away from the truth and can thus be updated more strongly. This result stands in contrast to previous findings on motivated reasoning (e.g., Lewandowsky et al., 2012; Jerit and Zhao, 2020), whereby preexisting worldviews or attachments to a political party can impede efforts to debunk fake news. However, Pennycook et al. (2020, 2021) argue that it is often a lack of attention rather than partisanship that drives such results. Our evidence is consistent with the latter line of thought (see also Section 5), since both the fact-checking and the media literacy intervention increase the awareness of fake news on Corona vaccines, and thereby induce participants to update their beliefs.

4.3. Persuasion rates

Sections 4.1 and 4.2 reveal that many participants are willing to get vaccinated or boostered against Covid-19 irrespective of the interventions. As a result, the ITT effects are relatively small. To adjust our

Table 7
Heterogeneity in baseline beliefs on Corona vaccination – MediaLiteracy.

	Wave I			Wave II		
Panel A: Media literacy – AfD supporters						
	Cred.	Dist.	Vaccine	Cred.	Dist.	Vaccine
	(1)	(2)	(3)	(4)	(5)	(6)
Media literacy	−0.134	−0.279	0.149	−0.161	0.090	0.136
	[0.096]	[0.166]	[0.071]	[0.091]	[0.170]	[0.068]
p-value	(0.167)	(0.095)	(0.039)	(0.080)	(0.600)	(0.047)
Controls	yes	yes	yes +	yes	yes	yes +
N	116	116	116	101	101	101
Baseline						
Mean DV	0.508	0.509	0.390	0.788	0.358	0.404
Std.Dev. DV	0.504	0.847	0.492	0.412	0.795	0.495
Panel B: Media literacy – non-AfD supporters						
	Wave I			Wave II		
	Cred.	Dist.	Vaccine	Cred.	Dist.	Vaccine
	(1)	(2)	(3)	(4)	(5)	(6)
Media literacy	−0.096	−0.208	0.021	−0.027	−0.151	0.042
	[0.024]	[0.050]	[0.020]	[0.031]	[0.056]	[0.023]
p-value	(0.000)	(0.000)	(0.280)	(0.379)	(0.007)	(0.075)
Controls	yes	yes	yes +	yes	yes	yes +
N	1,115	1,115	1,115	919	919	919
Baseline						
Mean DV	0.277	0.386	0.825	0.369	0.331	0.779
Std.Dev. DV	0.448	0.840	0.381	0.483	0.850	0.415

Notes: Table 7 displays the effect heterogeneity between AfD supporters (Panel A) and non-AfD supporters (Panel B) for our Media literacy intervention. In columns 1 and 4, the dependent variable is a dummy equal to one if participant *i* perceives the fakes on Corona vaccines as *Very credible*, *Credible* or *Indecisive* on average. In columns 2 and 5, the dependent variable is equal to participant *i*'s average standardized distance to the correct answer. In columns 3 and 6, the dependent variable is a dummy equal to one if participant *i* states to be *Very likely*, *Likely* or *Indecisive* to get vaccinated or boostered against Covid-19. All estimates are OLS estimates. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition. In columns 3 and 6 ("yes +"), we also control for participants' Corona vaccination status.

estimates for the share of participants left to be convinced, we compute persuasion rates à la DellaVigna and Kaplan (2007), both for the full sample as well as for AfD and non-AfD supporters, respectively. To that end, we define the share of participants left to be convinced as the NoIntervention group's share of participants stating to be *Very unlikely* or *Unlikely* to get vaccinated or boostered on the 5-point Likert scale.[48] Then, we divide the ITT estimates from our preferred specifications ("yes +") in Tables 4, 6, and 7 by that share.

Table 8 shows that around 21.7% (25.9%) of our participants could still be persuaded in Wave I (Wave II) of the survey. Specifically, 61% of the AfD supporters and 17.5% of the non-AfD supporters could still be persuaded in Wave I, and 59.6% (22.1%) in Wave II. Hence, the fact-checking intervention could convince about 7.8% (13.1%) of all persuadable participants in Wave I (Wave II) of the survey, whereas the media literacy intervention could convince 15.7% in Wave I and 18.5% in Wave II. These magnitudes are similar to those in comparable papers (e.g., Barrera et al., 2020, p.13). The persuasion rates for AfD supporters are considerably larger than the persuasion rates for non-AfD supporters. Hence, even if we account for differences in the share of participants left to be convinced, both interventions are more effective for AfD than for non-AfD supporters.[49]

A complementary explanation for our finding could be that it is *ceteris paribus* easier to convince the former. E.g., DellaVigna and Gentzkow (2010) argue that persuasion is more effective when receivers of novel information are less certain about the truth (p.654). Similarly, Kuklinski et al. (2000) distinguish between *mis*informed – those who have wrong beliefs and hold them firmly (p.792) – and *un*informed citizens. Consistent with that, we find that AfD supporters are on average less certain about their prior knowledge on current events, health, and nutrition. In particular, when we regress the uncertainty indicator from Section 3.4 on the AfD dummy, the resulting estimate indicates that AfD supporters are on average 4.1 percentage points more likely to be uncertain about their prior knowledge than non-AfD supporters; the effect is statistically significant at the 5%-level.[50] Similarly, when we further divide the sample of AfD supporters into those who are uncertain about their prior knowledge and those who are not, we find that the ITT estimates are larger and more statistically significant for the former group. Hence, the effect heterogeneity between AfD and non-AfD supporters is likely to be driven both by wrong priors and by uncertainty about them. Our results thus show that it is important not to give up on those with seemingly extreme opinions; rather, one should try to address exactly those people.

5. Mechanisms

Section 4.1 shows that the effectiveness of fact-checking tends to be limited to the fakes that are being corrected, while the media literacy intervention helps to distinguish between fakes and facts more generally, both immediately and in the short-run. A reasonable explanation is that the media literacy intervention raises participants' attention

[48] We obtain similar results using a cutoff here we define the share of participants left to be convinced as the NoIntervention group's share of participants stating to be *Very unlikely*, *Unlikely*, or *Indecisive* to get vaccinated or boostered.

[49] See Tables A.14 - A.12 in Appendix D for persuasion rates on our other binary outcomes.

[50] In contrast to that, there is no evidence that AfD supporters' prior knowledge on current events, health, and nutrition is worse than that of non-AfD supporters. The proportions of AfD and non-AfD supporters who are on the edge of being persuaded (i.e., stating to be *Undecided*) to get vaccinated or boostered are similar, too.

Table 8
Persuasion rates Corona vaccination.

Panel A: Fact-checking						
	Wave I			**Wave II**		
	Full	AfD	Non-AfD	Full	AfD	Non-AfD
	(1)	(2)	(3)	(4)	(5)	(6)
Persuasion rate	0.078	0.225	0.034	0.131	0.151	0.136
Share to be persuaded	0.217	0.610	0.175	0.259	0.596	0.221
Panel B: Media literacy						
	Wave I			**Wave II**		
	Full	AfD	Non-AfD	Full	AfD	Non-AfD
	(1)	(2)	(3)	(4)	(5)	(6)
Persuasion rate	0.157	0.244	0.120	0.185	0.228	0.190
Share to be persuaded	0.217	0.610	0.175	0.259	0.596	0.221

Notes: Table 8 displays the persuasion rates for our fact-checking (Panel A) and media literacy interventions (Panel B) for Wave I and Wave II of the survey, respectively. Columns 1 and 4 consider all participants in the NoIntervention, FactChecking, and MediaLiteracy groups. Columns 2 and 5 consider only AfD supporters, columns 3 and 6 only non-AfD supporters. The *Share to be persuaded* corresponds to the proportion of participants in the NoIntervention group stating to be *Very unlikely* or *Unlikely* to get vaccinated or boostered.

towards the Facebook postings and helps them to critically evaluate the postings' accuracy. Fact-checking, in contrast, turns participants into passive recipients of the specific corrections and thus fails to markedly enhance their skills.

To support the plausibility of this mechanism, this section shows that participants from the MediaLiteracy group are on average more likely to actively search for further information than participants from the NoIntervention group. In addition, they become better at considering untrustworthy elements in fakes and trustworthy elements in facts. Figures A.11 to A.13 in Appendix C illustrate our results; further details are given below. In Appendix E.3.6, we present results from a follow-up survey demonstrating that participants from the MediaLiteracy group spent significantly more time with the fakes and facts than participants from the NoIntervention group. Participants from the FactChecking group, in contrast, spent significantly less time with the postings. We interpret this as further evidence that the media literacy intervention raises awareness and critical evaluation, while fact-checking does not achieve this effect.

5.1. Search for further information

We first examine if participants actively search for further information. To this end, we generate a dummy equal to one if participant i reports to have used the Internet to answer the factual knowledge questions on Corona vaccines and nutrition, and use this dummy as dependent variable in Eq. (1).[51]

Consistent with the proposed mechanism, Table 9 shows that all estimates for the FactChecking group are negative, but they are not statistically significant (Panel A). In contrast to that, all estimates for the MediaLiteracy group are positive, and they are statistically significant at the 10%-level in Wave I of the survey (Panel B). Specifically, participants from the MediaLiteracy group are 4.7 to 4.9 percentage points more likely to search for further information than participants from NoIntervention group. The effect size corresponds to 9.8% of a standard deviation in the dependent variable in the baseline NoIntervention group.

5.2. Considering (un-)trustworthy elements in fakes and facts

Next, we show that the media literacy intervention helps participants to consider untrustworthy elements in fakes and trustworthy elements in facts, whereby they can better distinguish between false and correct information that they encounter online. To that end, we separately compute how many elements (in absolute terms) participants consider as trustworthy and untrustworthy, both in fakes and in facts.[52] For all of these measures we compare participants from the NoIntervention to the FactChecking and to the MediaLiteracy group.[53]

Table 10 confirms that the media literacy intervention induces participants to consider a larger absolute number of elements in fakes as untrustworthy, while the fact-checking intervention has no such effect. In particular, all estimates for fact-checking are close to zero and statistically insignificant (Panel A), while the estimates for the media literacy intervention are positive and statistically significant at the 1%-level (Panel B). In Wave I of the survey, participants from the MediaLiteracy group consider on average 1.2 more elements in fakes on Corona vaccines and 0.7 more elements in fakes on nutrition as untrustworthy than participants from the NoIntervention group.[54] For fakes on Corona vaccines, the effect size corresponds to 36.3% of a standard deviation in the dependent variable in the baseline NoIntervention group and to 36.7% of its mean value. For fakes on nutrition, the effect size corresponds to 38.5% of a standard deviation in the dependent variable in the baseline NoIntervention group and to 52.5% of its mean value.[55] The intervention's impact persists in Wave II of the survey. Specifically, participants from the MediaLiteracy group consider on average 0.9 more elements in fakes on Corona vaccines and 0.7 more elements in fakes on nutrition as untrustworthy than participants from the NoIntervention group. For fakes on Corona vaccines, this corresponds to 21.7% of a standard deviation in the dependent variable and to 20.1% of its mean value. For fakes on nutrition, this corresponds to 29.1% of a standard deviation in the dependent variable and to 41.2% of its mean value.

Analogously, Table 11 confirms that the media literacy intervention induces participants to consider a larger number of elements in facts as trustworthy, while the fact-checking intervention has no such effect. In particular, all estimates for fact-checking are close to zero and statistically insignificant (Panel A), while the estimates for the media literacy intervention are positive and statistically significant at the 1%-level (Panel B). In Wave I of the survey, participants from the MediaLiteracy group consider on average 1 more element in facts on Corona vaccines and 0.7 more elements in facts on nutrition as trustworthy than participants from the NoIntervention group. For facts on Corona vaccines, the

[51] We did not pre-register the question on search for further information.

[52] For more details on how the measure is constructed see Section 3.4, a detailed description of how and which elements can be clicked is available in Section 3.1.1.

[53] We did not pre-register the consideration of (un-)trustworthy elements.

[54] Table A.17 in Appendix D reports exemplary results for one specific type of elements: *Verified account*. In line with the results from this section, we find that participants from the MediaLiteracy group are more likely to consider a *Verified account* label in facts on Covid-19 and nutrition as trustworthy than participants in the NoIntervention group. The fact-checking intervention, in contrast, has no such effect.

[55] All differences between the FactChecking and the MediaLiteracy group are statistically significant at the 1%-level (two-sided t-tests, $p < 0.000$).

Table 9
Search for further information.

Panel A: Fact-checking								
	Wave I				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fact-checking	−0.041	−0.034	−0.037	−0.035	−0.015	−0.008	−0.032	−0.026
	[0.028]	[0.027]	[0.025]	[0.025]	[0.030]	[0.030]	[0.028]	[0.028]
p-value	(0.144)	(0.208)	(0.150)	(0.165)	(0.628)	(0.777)	(0.261)	(0.338)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,225	1,225	1,225	1,225	1,022	1,022	1,022	1,022
Panel B: Media literacy								
	Wave I				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Media literacy	0.049	0.048	0.047	0.047	0.010	0.016	0.036	0.043
	[0.028]	[0.028]	[0.027]	[0.026]	[0.030]	[0.030]	[0.029]	[0.029]
p-value	(0.083)	(0.085)	(0.083)	(0.079)	(0.733)	(0.586)	(0.215)	(0.136)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,231	1,231	1,231	1,231	1,020	1,020	1,020	1,020
Baseline: No Intervention								
Mean DV	0.400	0.400	0.307	0.307	0.361	0.361	0.295	0.295
Std.Dev. DV	0.490	0.490	0.462	0.462	0.481	0.481	0.456	0.456

Notes: Table 9 shows the OLS estimates of comparing the NoIntervention to the FactChecking (Panel A) and to the MediaLiteracy group (Panel B), respectively. The dependent variable is a dummy equal to one if participant *i* reports to have used the Internet to respond to the factual knowledge questions (fakes and facts) on Corona vaccines and nutrition in Wave I and in Wave II, respectively. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition.

effect corresponds to 33.5% of a standard deviation in the respective dependent variable in the baseline NoIntervention group and to 38.3% of its mean value. For facts on nutrition, the effect corresponds to 20.9% of a standard deviation in the respective dependent variable in the baseline NoIntervention group and to 20.7% of its mean value. Again, the impact of the media literacy intervention persists in Wave II of the survey. Specifically, participants from the MediaLiteracy group consider on average 0.8 more elements in facts on Corona vaccines and 0.7 more elements in facts on nutrition as trustworthy than participants from the NoIntervention group. For facts on Corona vaccines, this corresponds to 38.5% of a standard deviation in the dependent variable and to 36.3% of its mean value. For facts on nutrition, this corresponds to 25.3% of a standard deviation in the dependent variable and to 22.8% of its mean value.

Crucially, participants from the MediaLiteracy group do not generally consider a larger number of elements as (un-)trustworthy in the fakes and facts. In particular, the media literacy intervention does not increase the number of elements considered as trustworthy in fakes (Table A.15 in Appendix D), and its impact on the number of elements considered as untrustworthy in facts is small and limited to (arguably untrustworthy) emojis in facts on nutrition (Table A.16 in Appendix D). Hence, consistent with the results from Section 4.1, the media literacy intervention does not make participants more skeptical towards social media postings per se, but rather helps them to distinguish between trustworthy and untrustworthy (elements of the) information.[56]

6. Further analyses

This section provides further context for the interpretation of our main results. Specifically, we examine the PassiveControl group to study if our interventions can offset the damage that fake news are causing, we consider the JustFacts group to further investigate the limited effectiveness of fact-checking, we discuss potential heterogeneity in the fakes and fact-checks that we show to our participants, and we provide a battery of robustness checks for our main results.[57]

6.1. Comparison to PassiveControl group

The main purpose of our paper is to study whether and to what extent fact-checking and media literacy interventions are able to debunk fake news that circulate on social media. The relevant benchmark are thus participants from the NoIntervention group, who are exposed to fakes and facts without further intervention. While this comparison illustrates the effectiveness of our interventions in an environment where all participants see fakes and facts, it does not uncover to what extent the interventions are able to reverse the harm the fakes are causing. To better interpret the magnitude of our main estimates in that regard, this section compares responses from the PassiveControl group – i.e., participants who did not see any fakes or facts from Facebook at all – to the FactChecking and the MediaLiteracy group, respectively. The smaller the difference between those groups, the more effective is the respective intervention in repealing the impact of fakes.[58] We complement the analysis with a comparison of the PassiveControl to the NoIntervention group, which provides a benchmark for the absolute impact of fakes. Figures A.9 and A.10 in Appendix C illustrate our results; further details are provided below.[59]

6.1.1. Factual knowledge

Table A.18 in Appendix D reveals that exposure to fakes on Corona vaccines and nutrition substantially impairs participants' factual knowledge, and that neither the fact-checking nor the media literacy intervention can fully offset the effect. In particular, we find that (almost) all estimates are positive and statistically significant at the 1%-level, which means that responses by the FactChecking, the MediaLiteracy, and

[56] Note that our primary objective is to assess whether participants in the MediaLiteracy group improve in distinguishing trustworthy and untrustworthy elements within a post, rather than evaluating the post's overall accuracy, which likely depends on the interplay of all elements.

[57] We preregistered our analyses with a PassiveControl and a JustFacts group; see Appendix B for details.

[58] Our interventions do not overcompensate the impact of fakes.

[59] Recall that participants in the PassiveControl group are not asked to rate the credibility of fakes and facts (see Section 3.1), since the idea of the PassiveControl group is to measure factual knowledge and attitudes when not being exposed to any of the fakes or facts. Therefore, we cannot compare the credibility outcomes measured in the main treatments with the PassiveControl group.

Table 10
Absolute number of untrustworthy elements considered in fakes.

Panel A: Fact-checking								
	Wave I				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fact-checking	−0.023	−0.093	−0.048	−0.085	0.070	−0.022	0.014	−0.041
	[0.190]	[0.184]	[0.099]	[0.097]	[0.264]	[0.256]	[0.138]	[0.135]
p-value	(0.902)	(0.610)	(0.627)	(0.383)	(0.792)	(0.931)	(0.917)	(0.759)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,225	1,225	1,225	1,225	1,030	1,030	1,030	1,030
Panel B: Media literacy								
	Wave I				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Media literacy	1.242	1.194	0.679	0.660	1.088	0.903	0.725	0.680
	[0.210]	[0.197]	[0.109]	[0.106]	[0.290]	[0.271]	[0.163]	[0.155]
p-value	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,231	1,231	1,231	1,231	1,026	1,026	1,026	1,026
Baseline: No Intervention								
Mean DV	3.249	3.249	1.256	1.256	4.469	4.469	1.651	1.651
Std.Dev. DV	3.293	3.293	1.714	1.714	4.144	4.144	2.337	2.337

Notes: Table 10 compares the absolute number of elements considered as untrustworthy in fakes on Corona vaccines and nutrition for participants from the NoIntervention to the FactChecking (Panel A) and the MediaLiteracy group (Panel B) in Wave I and Wave II of the survey, respectively. All estimates are OLS estimates. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition.

Table 11
Absolute number of trustworthy elements considered in facts.

Panel A: Fact-checking								
	Wave I				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fact-checking	0.075	0.026	0.011	0.009	0.042	0.023	0.077	0.051
	[0.167]	[0.162]	[0.187]	[0.184]	[0.138]	[0.136]	[0.180]	[0.181]
p-value	(0.653)	(0.872)	(0.953)	(0.962)	(0.763)	(0.867)	(0.670)	(0.777)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,225	1,225	1,225	1,225	1,030	1,030	1,030	1,030
Panel B: Media literacy								
	Wave I				**Wave II**			
	Corona		Nutrition		Corona		Nutrition	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Media literacy	1.059	0.998	0.741	0.672	0.873	0.809	0.758	0.686
	[0.196]	[0.185]	[0.200]	[0.193]	[0.158]	[0.149]	[0.189]	[0.182]
p-value	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Controls	no	yes	no	yes	no	yes	no	yes
N	1,231	1,231	1,231	1,231	1,026	1,026	1,026	1,026
Baseline: No Intervention								
Mean DV	2.607	2.607	3.239	3.239	2.231	2.231	3.012	3.012
Std.Dev. DV	2.975	2.975	3.205	3.205	2.103	2.103	2.711	2.711

Notes: Table 11 compares the absolute number of elements considered as trustworthy in facts on Corona vaccines and nutrition for participants from the NoIntervention to the FactChecking (Panel A) and the MediaLiteracy group (Panel B) in Wave I and Wave II of the survey, respectively. All estimates are OLS estimates. Robust standard errors in squared parentheses, p-values in round parentheses. Control variables include age, gender, family status, household earnings, education, personality traits ("big 5"), political preferences, and prior knowledge on current events, health, and nutrition.

the NoIntervention group are significantly further away from the correct answer than responses by the PassiveControl group.

There are two additional insights. First, in comparison to the absolute impact of fakes (Panel C), the average effectiveness of our interventions is relatively small. E.g., in Wave I of the survey, the fact-checking intervention repeals less than 40% of the damage caused by fakes on Corona vaccines, and 80% of the damage caused by fakes on nutrition. Similarly, the media literacy intervention repeals just 26.2% of the damage caused by fakes on Corona vaccines, and 80% for nutrition. Thus, while both interventions improve factual knowledge relative to the NoIntervention group – i.e., in an environment where all participants see fakes and facts – they do not reverse the harm of fakes entirely.

Second, the estimates for nutrition are smaller than the estimates for Corona vaccines, i.e., responses are on average more similar to the PassiveControl group. A plausible explanation could be that the fakes on Corona vaccines are more persuasive than the fakes on nutrition, whereby participants' factual knowledge on the former is shifted further away from their prior than their factual knowledge on the latter.

6.1.2. Attitudes

In contrast to factual knowledge, Table A.19 in Appendix D shows that exposure to fakes has a relatively small, if any, impact on participants' attitudes (Panel C), and that both the fact-checking and, in particular, the media literacy intervention can effectively repeal that

impact. Specifically, we find that (almost) all estimates are close to zero and statistically insignificant, which means that participants from the FactChecking, the MediaLiteracy, and the NoIntervention group report a similar willingness to get vaccinated (or boostered) against Covid-19 as well as a similar willingness to consume (needless) dietary supplements as participants from the PassiveControl group. This confirms the results from Section 4, whereby the majority of participants wants to get vaccinated or boostered irrespective of the interventions, and whereby it is difficult to affect habit-based attitudes on dietary supplements. Hence, the modest impact of our interventions on participants' attitudes (see Table 4) can also be explained by the small absolute impact of fakes: if exposure to fakes does not sway participants' attitudes to begin with, there is nothing that the fact-checking or the media literacy intervention could change.[60]

6.2. Comparison to JustFacts group

Section 5 demonstrates that the media literacy intervention helps participants to better distinguish between fakes and facts, while fact-checking fails to markedly enhance their skills. A complementary explanation for the smaller effectiveness of fact-checking is that the corrections often repeat false claims and thus induce "anchoring" (Tversky and Kahneman, 1974) or "continued influence effects" (Lewandowsky et al., 2012), whereby users' beliefs are biased towards the initially presented values. To study the role of such effects in our context, this Section compares factual knowledge of the JustFacts group – i.e., participants who only saw facts and fact-checks – to the NoIntervention and the PassiveControl group, respectively. If anchoring effects do not play a role, participants from the JustFacts group should have better factual knowledge than participants from the NoIntervention and the PassiveControl group, because they are given the correct information. If anchoring effects exist, however, we expect the JustFacts group's factual knowledge to be in between the NoIntervention and the PassiveControl group. The closer their responses are to the former, the stronger are the anchoring effects. Figure A.9 in Appendix C illustrates our results, further details are discussed below.

Table A.20 in Appendix D reveals that participants from the JustFacts have better factual knowledge on nutrition than participants from the PassiveControl, and better factual knowledge on both Corona vaccines and nutrition than participants from the NoIntervention group.[61] All differences are statistically significant at the 1%- or at the 5%-level.

There are three potential explanations for these results that are not mutually exclusive. First, consistent with Tversky and Kahneman (1974) and Lewandowsky et al. (2012), repetition of the false claims could stick in participants' memory. In particular, while all fact-checks on Corona vaccines restate the respective fake news, the fact-checks on nutrition do not recast any false or misleading numbers. As a result, fact-checking increases participants' factual knowledge on nutrition (as measured by the PassiveControl group), but reduces factual knowledge on Corona vaccines (although the JustFacts group still performs significantly better than the NoIntervention group).[62] Second, Section 6.4 reveals that participants from the JustFacts group perceive the fact-checks on nutrition as significantly more credible than the fact-checks on Corona vaccines. Thus, it could be that the former exhibit a stronger impact on participants' priors, whereby they update their beliefs more extensively. Third and relatedly, participants' prior beliefs on nutrition could be less firm than their beliefs on Corona vaccines, and thereby more easy to sway. Similarly, as discussed in Section 6.1 above, their priors on nutrition could be worse than their priors on Corona vaccines (i.e., further away from the truth), leaving more room for improvement through the fact-checks. In sum, our evidence is in line with "anchoring" or "continued influence effects" that constitute one potential drawback of fact-checking, but we cannot exclude alternative explanations, either.

6.3. Heterogeneity in credibility of fakes

The effectiveness of our interventions is likely to depend on the specific fakes that we select for the experiment. To explore this issue in more detail, Figure A.14 in Appendix C displays the (disaggregated) mean credibility of all fakes as given by the NoIntervention group on a 5-point Likert Scale.

We find that there is substantial heterogeneity in the credibility of fakes. In particular, participants perceive the fakes on Corona vaccines on average as less credible than the fakes on nutrition. Moreover, there is heterogeneity within topics: the perceived credibility of fakes on Corona vaccines ranges from 1.65 to 2.28 and of fakes on nutrition from 2.75 to 3.41.

To further examine how fake credibility affects intervention effectiveness, we conduct a follow-up experiment where the NoIntervention group rates the fakes encountered during the first part of the survey as more credible than in our original study. Interestingly, in contrast to our original study, we detect a small but statistically significant effect of fact-checking on non-fact-checked posts in this follow-up. Fact-checks targeting less obvious fakes may thus lead readers to approach subsequent posts with greater criticism. Moreover, the findings of the follow-up strongly support our interpretation that the media literacy intervention enhances skills that can be more broadly applied than fact-checking. A detailed description of the follow-up is provided in Appendix E. For a direct comparison of intervention effects on nutrition posts in the original and follow-up experiments, see Table A.29 in Appendix E.

6.4. Heterogeneity in credibility of fact-checks

According to Jerit and Zhao (2020), trust in the authors of corrective messages is a crucial cause for their effectiveness. In our context, distrust in the (authors of the) fact-checks could further explain why the fact-checking is less effective than the media literacy intervention. To explore the plausibility of this explanation, Figure A.15 in Appendix C displays the (disaggregated) mean credibility of all fact-checks on a 5-point Likert Scale.[63]

We find that the mean credibility for fact-checks on Corona vaccines in Wave I of the survey – i.e., those that were displayed to the FactChecking group – is surprisingly low: the fact-checks rate between 2.62 and 2.73 for participants of the FactChecking, and between 2.81 and 2.88 for participants of the JustFacts group. This is significantly less than for fact-checks on Corona vaccines in Wave II of the survey (ratings between 3.15 and 3.27) and for fact-checks on nutrition in either Wave (ratings between 3.60 and 3.79). One possible driver of these differences could be heterogeneity in the source. E.g., while both fact-checks on Corona vaccines in Wave II of the survey are released by *dpa*, Germany's most renowned news wire, the fact-checks on Corona vaccines in Wave I stem from *Correctiv* and *AFP*, respectively. Although these are generally considered as reliable fact-checking organizations,

[60] This is in contrast to Barrera et al. (2020), who find that exposure to fake news is highly persuasive. However, Barrera et al. (2020) consider fake news on migration in France, while we consider fake news on Corona vaccines and nutrition. The diverging results could thus be driven by differences in the context of the fakes, i.e., it could be easier to sway participants' voting intentions than their attitudes on Corona vaccination and dietary supplements.

[61] Due to a technical issue with one of the fact-checks on nutrition in Wave I of the survey, we do not aggregate participants' responses to the factual knowledge questions but consider just the one functioning fact-check instead.

[62] This explanation would also be consistent with the salience effects documented by Barrera et al. (2020).

[63] Recall that the FactChecking group was only shown fact-checks on Corona vaccines in Wave I, while the JustFacts group saw fact-checks on all topics in both waves of the survey.

they might be less known among the participants of our experiment, and thus perceived as less trustworthy. On the other hand, one of the fact-checks on nutrition in Wave II of the survey comes "just" from an online platform and three from national public authorities, but Figure A.15 shows that there are just minor differences in their perceived credibility. Hence, while it seems plausible that small trust in fact-checking contributes to its relative ineffectiveness, we cannot claim with certainty that the authors of the fact-checks are crucial components in this.

How does this affect the interpretation of our results? While the existing literature agrees that trust in fact-checks is a crucial cause for their effectiveness, it also shows that little trust in fact-checks does not necessarily mean that fact-checking does not work at all: E.g., Martel and Rand (2024) find that while warning effects are somewhat smaller among participants with lower trust in the fact-checkers, the warning labels still significantly reduced belief in false information (by 12.9%, p.1960), even among those most skeptical of fact-checkers. Similarly, Liu et al. (2023) observed few differences in effectiveness across various fact-checking sources, although sources deemed more credible showed greater effectiveness.

These findings suggest that the effects we observe regarding the efficacy of fact-checking on the perceived credibility of fakes on Covid-19 vaccines likely represent lower bounds. The effects could potentially be even stronger if trust in the fact-checks was higher.

6.5. Order of fact-checks

In our main experiment fact-checks were displayed to the participants before seeing the actual fake post. In reality however, platforms implement fact-checks both before and after posts with fake content. We randomized this order in the follow-up experiment to test whether the order of fact-check and fake influences the initial experiment's findings. A comparison of the effectiveness of fact-checking for the different orders within the FactChecking group suggests weak evidence that fact-checking is more effective when fact-checks are presented after the fake. However, this result is not robust to a slightly stricter cutoff in the dependent variable (while the findings in our original experiment are robust to different cutoffs).

As we observe only small and not very robust differences for the order of the fact-check, we conclude that the influence of the order is most likely not the driving force behind our main results. If the order of the fact-check played a role in our main experiment, the estimates we obtained there should be regarded as lower bounds. Moreover, the qualitative nature of our findings with respect to the MediaLiteracy group remains unchanged: Even when fact-checks are displayed after the fake the effect of the fact-checking intervention is substantially lower than that of the media literacy intervention. A detailed description of this analysis and regression results are denoted in Appendix E.3.4.

6.6. Robustness checks

This section provides several analyses that support the robustness of our main results. In particular, we show that heterogeneity in terms of participants' vaccination status and the time span between Waves I and II does not play a role, we present evidence from list experiments that support the validity of the participants' self-reported attitudes, and we demonstrate that our results are robust to using an IV approach, where we take potential non-compliance into account.

6.6.1. Heterogeneity in terms of vaccination status

At the time of our experiment (September/October 2021), every German adult could be fully vaccinated against Covid-19 (two injections); yet, only about 81.8% of our participants reported to have taken the opportunity. This raises two potential concerns. First, participants who selected themselves into vaccination could be systematically different from those who did not, especially with respect to the questions on Corona vaccines. Second, we elicited participants' attitudes towards Corona vaccination with two different questions – "willingness to get vaccinated" vs. "willingness to get boostered" on a 5-point Likert-scale, respectively (see Section 3) – and responses to those two questions might not be entirely comparable. Our main analyses address these concerns with robustness checks, where we add participants' Corona vaccination status as an additional control (see Section 4). To further support our main findings, this section shows that they are robust to splitting the sample into participants who are fully vaccinated and those who are not.[64]

Consistent with the robustness checks that we already conducted, Tables A.23 and A.24 in Appendix D show that the effect of our fact-checking and media literacy interventions are similar for participants who are fully vaccinated and those who are not. The ITTs of fact-checking (Table A.23), for instance, hardly differ between the two subsamples; the main difference is the reduced precision in Panel B, stemming from the small sample of non-vaccinated participants. Table A.24 reveals that the effect on the willingness to get vaccinated is larger for non-vaccinated than fully vaccinated participants in the MediaLiteracy group. This result is comparable to our findings on effect heterogeneity for AfD and non-AfD supporters from Section 4.2. It likely arises from differences in the baseline (i.e., there are more non-vaccinated participants left to be convinced) and, to some degree, from the differently posed question. In sum, however, we conclude that heterogeneity between participants who are fully vaccinated and those who are not is at most a minor concern.

6.6.2. Delay between waves I and II

As we describe in Section 3.2, all participants received a re-invitation to Wave II about one week after they completed Wave I of the survey. However, as they could re-start the survey at any time after that, the delay between actual participation in the two waves is quite heterogeneous and lies between 8 and 45 days.[65]

We conduct two analyses to confirm that heterogeneity in the delay between Waves I and II is unlikely to affect our results. First, when we regress the number of days between participation in the two waves on our treatment indicators (with the NoIntervention group as omitted category) plus the full set of controls, almost all estimates are close to zero and statistically insignificant.[66] Hence, our treatments do not seem to influence when participants re-start the survey. Second, there is no evidence that differences in the delay between Waves I and II affect the magnitude of our main estimates. When we interact the number of days between participation in the two waves with the treatment indicators in Eq. (1), almost all interaction terms are close to zero and statistically insignificant. Similarly, when we split the sample at the median delay (= 15 days) and estimate Eq. (1) on the two subsamples, respectively, the resulting point estimates resemble those from Section 4.1. Note, however, that the delay between actual participation in Wave I and II is endogenous, so all robustness checks from this section should be interpreted with caution.

To interpret these results, note that a two-week delay is relatively short compared to other education intervention studies, where effects often persist for months (see Kaiser et al., 2022, for an overview). Thus, it is unsurprising that the effects are similar for participants who completed the survey at just slightly different times. Our findings also align with the literature on correcting misinformation. For instance, Guess et al. (2020) found that the effects of their media literacy intervention remained measurable and statistically significant after several weeks.

[64] For brevity, we only report the results for questions related to Corona vaccines. We do not find any systematic effect heterogeneity between participants who are fully vaccinated and those who are not with respect to questions on nutrition.

[65] We did not preregister the analysis of the delay between Waves I and II.

[66] The estimate for the PassiveControl group is negative and weakly statistically significant at the 10%-level, indicating that participants re-started the survey about half a day earlier than participants from the NoIntervention group.

Table 12
List experiments.

	All participants				No Intervention				Fact-checking				Media Literacy			
	Wave I		**Wave II**		**Wave I**		**Wave II**		**Wave I**		**Wave II**		**Wave I**		**Wave II**	
	Vacc. (1)	Suppl. (2)	Vacc. (3)	Suppl. (4)	Vacc. (5)	Suppl. (6)	Vacc. (7)	Suppl. (8)	Vacc. (9)	Suppl. (10)	Vacc. (11)	Suppl. (12)	Vacc. (13)	Suppl. (14)	Vacc. (15)	Suppl. (16)
Direct question	0.19	0.30	0.22	0.31	0.22	0.30	0.26	0.30	0.17	0.29	0.20	0.31	0.16	0.27	0.20	0.28
List experiment	0.16	0.38	0.08	0.33	0.21	0.34	0.11	0.31	0.14	0.37	0.07	0.45	0.06	0.40	0.01	0.26
N	3,051	3,051	2,525	2,525	618	618	509	509	607	607	513	513	613	613	511	511

Notes: In row 1, Table 12 displays the proportion of participants who in the main experiment **directly** report to be *Very unlikely* or *Unlikely* to get vaccinated or boostered against Covid-19 (columns 1 and 3) and who **directly** report to be *Very likely* or *Likely* to consume dietary supplements (columns 2 and 4). In row 2, Table 12 displays the respective indirectly elicited proportions from the list experiments.

Similarly, Carey et al. (2022) studied the decay in fact-checking efficacy and, consistent with our results, found no effect approximately two weeks post-intervention.

6.6.3. List experiments and experimenter demand bias

Participants' self-reported attitudes on Corona vaccines and dietary supplements might suffer from social desirability or experimenter demand bias if the participants anticipate that we as researchers are in favor of vaccination and against the consumption of needless dietary supplements. However, recent empirical evidence from both the behavioral economics as well as the political science literature suggests that experimenter demand biases are small, if they exist at all (De Quidt et al., 2018; Mummolo and Peterson, 2019). As argued in Section 3, we further minimize the risk of experimenter demand bias by *not* incentivizing the corresponding questions with a potential bonus payment. In addition to that, this section conducts two list experiments à la Blair and Imai (2012) – one for Corona vaccines, one for nutrition – to confirm that our participants do not conceal any socially undesirable opinions and attitudes.

For each list experiment, we randomly partition our participants into two groups. One group receives a list of five, the other group a list of six statements in random order, where the additional sixth statement is "I prefer not to get vaccinated against Covid-19." ("I take dietary supplements.") and the other five statements are about unrelated topics (see Appendix A.2 for the full lists). Then, we ask each participant how many of those statements he or she would agree with. Finally, we compute the difference in means between the two groups for the number of supported statements, which can be interpreted as the proportion of participants who indirectly concede that they do not want to get vaccinated (that they consume dietary supplements). If these proportions are substantially larger than the proportions that we directly elicit in the experiment, the self-reported attitudes might suffer from social desirability or experimenter demand bias.

Columns 1 to 4 in Table 12 show that the proportion of participants who directly report that they do not want to get vaccinated or boostered is similar (column 1) or even larger (column 3) than the proportion elicited through the list experiment. The proportions of participants who directly report to consume dietary supplements are smaller than the proportion that we elicit through the list experiment (columns 2 and 4), but the differences are small and could be driven by the slightly different questions in the main and in the list experiment ("How likely are you to consume dietary supplements in the near future?" vs. "I consume dietary supplements.") In sum, there is no evidence for systematic experimenter demand bias with respect to the self-reported attitudes on Corona vaccination and dietary supplements when we consider the entire sample.[67]

Next, we check for experimenter demand bias within our three main treatment groups. Given our experimental setup, experimenter demand bias is especially likely to occur for the FactChecking group when we ask about their willingness to get vaccinated in Wave I of the survey, and for the MediaLiteracy group throughout. However, as for the entire sample, the proportions of participants who directly state that they do not want to get vaccinated are even larger than the proportions of participants who indirectly say so, both for the FactChecking and for the MediaLiteracy group, and in both Wave I and Wave II of the survey. Hence, we do not find evidence for experimenter demand bias here. We do observe that for the MediaLiteracy group, the proportion of participants who just indirectly report to consume dietary supplements is larger than the proportion of participants who directly says so. However, this is also true for the FactChecking and the NoIntervention group, where experimenter demand bias is unlikely to occur, which suggests that the result is driven by unrelated issues. In sum, we can conclude that experimenter demand bias is unlikely to confound our main results.

6.6.4. IV analysis

Participants from the FactChecking and the MediaLiteracy group might skip the intervention by just quickly clicking through the survey. In this case, the ITT estimates would underestimate the average treatment effect. To take this into account, this section presents an IV approach, where we use participants' time spent with the interventions to determine their actual treatment status and their random assignment to a treatment group as an instrument.

We proceed in two steps. First, we specify how much time it takes to properly engage with the interventions. To this end, we asked eleven Research Assistants to carefully read the two fact-checks as well as the ten tips to spot false news and recorded how much time they need. We find that the minimum amount of time spent on the fact-checking intervention is equal to 24.7, and the minimum amount of time spent on the media literacy intervention is equal to 38.9 s.

Second, we define D_i as a dummy variable that indicates participant i's actual treatment status. In particular, D_i is equal to one if i spent at least 24.7 s with the fact-checking or 38.9 s with the media literacy intervention. Thus, D_i is equal to zero for participants who did not spent a reasonable amount of time with their respective intervention, and for all participants in the NoIntervention group.[68] Eq. (1) thus extends to

[67] We further support the analysis with a sample split (see Table A.21 in Appendix D), where we consider participants who directly report that they are likely to get vaccinated (consume dietary supplements) on the one hand, and participants who report that they are unlikely to get vaccinated (consume dietary supplements) on the other. Reassuringly, we find that the proportion of participants who indirectly concede that they are not going to get vaccinated against Covid-19 is much larger for the latter than for the former group. Similarly, the proportion of participants who indirectly concede to consume dietary supplements is much larger for participants who directly report that they are likely to do so than for participants who report that they are not.

[68] Using the minimum amount of time spent with the interventions as our threshold is the most conservative choice. When we use the median or mean amount of time from surveying the Research Assistants, the IV estimates become larger, but are qualitatively unaffected.

$$y_{iw} = \gamma_0 + \gamma_1 \widehat{D_i} + \gamma_2 X_i + \epsilon_i \quad (3)$$

$$D_i = \pi_0 + \pi_1 TG_i + \pi_2 X_i + u_i, \quad (4)$$

which we estimate by 2SLS.

We prefer using a binary (rather than a continuous) measure for participants' actual treatment status for two reasons. First, the impact of time spent with the interventions is likely to be discrete: participants need a certain minimum amount of time to understand and process the novel information, but any time spent beyond that is unlikely to yield further benefits. Second, it generally takes more time to engage with the media literacy than with the fact-checking intervention. Hence, using a binary measure for participants' actual treatment status makes the regression results better comparable across treatment groups.

Table A.22 in Appendix D confirms that the 2SLS estimates of Eqs. (3) and (4) are larger, but qualitatively similar to their counterparts from Section 4.1.[69] Moreover, the coefficients for π_1 demonstrate that close to 70% of the FactChecking, and 74% of the MediaLiteracy group spent a considerable amount of time with their respective intervention. Skipping the interventions could be a larger concern outside the context of our experiment, though, especially if they disrupt users' consumption of social media. We further discuss this issue below.

7. Conclusion

We conduct a large-scale randomized survey experiment on the immediate and short-term effects of fact-checking and media literacy interventions to demonstrate that the impact of fact-checking tends to be limited to the fakes that are being corrected, whereas the media literacy intervention helps to distinguish between fakes and facts more generally, both immediately and in the short-run. A plausible mechanism for our result is that the media literacy intervention enables participants to critically evaluate social media postings, while fact-checking turns them into passive recipients of the specific corrections and thus fails to markedly enhance their skills. Hence, in an environment where not every claim can be fact-checked, the media literacy intervention is likely to be more effective than fact-checking on average.

Our paper promotes brief media literacy interventions as an effective tool to fight fake news and advances current policy debates along these lines. The European Union, for instance, has recently asserted media literacy as a pivotal tool to counter misinformation on social media,[70] and the UNESCO has provided policy guidelines for digital media and information literacy.[71] Our results strongly support such endeavors and suggest that official media literacy campaigns – which are relatively cheap, scalable, and easy-to-implement – could be a valuable complement to existing efforts like fact-checking.

Media literacy interventions as a means to fight fake news might outperform fact-checking for three more reasons. First, many news items are not clearly fake or fact. While fact-checkers can only debunk fake news that clearly provide wrong or misleading information, media literacy interventions are likely to raise users' skills and awareness to spot even more subtle types of misinformation. Second, the effectiveness of fact-checking hinges on users' trust in the fact-checker. E.g., about 50% of Americans think that fact-checkers are biased (Allen et al., 2021). In times where users' trust in renowned sources like public service broadcasters is crumbling, the provision of fact-checks is likely to be futile, as those whom the intervention is supposed to target are often those who distrust the fact-checker. Third and relatedly, with the rise of Artificial Intelligence (AI), users might have to increasingly rely on their own critical judgment of what they see rather than blindly trust the assessment of third parties. This applies even when AI is being used by researchers and journalists to *assist* the detection of fake news. Specifically, while AI is currently able to detect suspicious and check-worthy news items and match them to claims that have previously been fact-checked, fact-checking as such is still based on manual assessment and far from being fully automated.[72]

Our analysis has several limitations that open avenues for further research. First, the magnitude of our coefficients is likely to depend on the specific fakes, facts, and fact-checks as well as on the topics that we selected for the experiment, and other details of the implementation such as the absence of placebo treatments. Therefore, we consider the qualitative results as our most insightful findings and recommend to interpret the precise point estimates with caution. E.g., some fakes are harder to detect than others, which is likely to reduce the effectiveness of our interventions. Also, there exist different types of fact-checking (e.g., journalist-based vs. crowd-sourced fact-checking, see Allen et al. (2021)), whose effectiveness might vary in different contexts and depend on the recipient. Similarly, users may be more or less well informed about different topics, whereby they are more or less likely to benefit from our interventions.

Our sample is representative of the German population in terms of age and gender, suggesting that our findings are generalizable to countries with similar demographics, such as many in the EU.[73] Notably, surveys suggest no heightened sensitivity among German users to fake news or media literacy initiatives. For example, evidence from *RavenPack* indicates that German news media discuss fake news related to Covid-19 less frequently than other EU countries.[74] Furthermore, Germans report encountering fake news less often than respondents from other countries.[75] While fake news about dietary supplements is prevalent on social media,[76] 56% of Germans feel well-informed about the associated risks and benefits.[77] Additionally, although Germany implemented the Network Enforcement Act in 2017, its impact on user behavior, particularly on Facebook, appears minimal (Maaß et al., 2024). While we cannot rule out that our results are specific for a German audience, the mentioned evidence lowers our concerns about a limited external validity beyond the German context – especially regarding the qualitative results.

Second, it is unclear how many and which users would actually engage with media literacy interventions. In particular, some users might perceive such trainings as a nuisance and consequently skip them. In addition, it could be that mostly users who are well informed anyway decide to take part in media literacy interventions, while users with poor priors – i.e., those for whom the intervention would be most effective – prefer to shirk them. Participation in media literacy interventions will ultimately depend on their design. However, even if such interventions fail to reach the entire population, it is worthwhile to enhance the skills even of a subset of users, and this should be preferred over not doing anything. Thus, while our paper stresses the

[69] As further robustness checks for the MediaLiteracy group, we replace D_i with a dummy that is (i) equal to one if participant i reports to have used the tips during the experiment and (ii) equal to one if participant i could recall the tips correctly. The 2SLS estimates are larger than their counterparts in Table A.22 in both cases, but qualitatively unaffected.

[70] See https://digital-strategy.ec.europa.eu/en/policies/media-literacy (viewed August 2022).

[71] See https://www.unesco.org/en/communication-information/media-information-literacy/policy-strategy (viewed August 2022).

[72] See EU Horizon (https://ec.europa.eu/research-and-innovation/en/horizon-magazine/can-artificial-intelligence-help-end-fake-news, viewed August 2023).

[73] See Eurostat: Demography of Europe -- 2023 (viewed December 2024).

[74] See https://coronavirus.ravenpack.com/ (viewed April 2024).

[75] See survey results by *Statista*: 1 , 2 (viewed December 2024).

[76] See media articles by Eisele (Deutsche Welle, viewed December 2024: *Faktencheck: Abzocke mit Nahrungsergänzungsmitteln?*) and Schaetze and Lerch (consumer protection agency, viewed December 2024: *Gesundheitsversprechen für Nahrungsergänzungsmittel auf Instagram – häufig abseits der Legalität*).

[77] See *Forsa-Survey* on *Opinion on Nutrition Supplements* available here *https://verbraucherzentrale.de/* (viewed December 2024).

high potential of media literacy interventions as a tool to fight fake news, the most efficient ways to implement such interventions must be examined in future research.

Third and relatedly, while we demonstrate that media literacy interventions could help users to better distinguish between false and correct information that they encounter online, we remain agnostic about the concrete implementation of such interventions on behalf of social media platforms. In particular, it is unclear if social media would be willing to set up regular interventions (e.g., in terms of pop-up windows that appear every few weeks) and what the ideal type of intervention would look like. The fact that Facebook has developed a set of "Tips to Spot False News" on its own behalf is encouraging, though, and suggests that social media might be willing to cooperate with academics and policy makers. The ideal type of intervention is likely to depend on the specific social media platform – e.g., users on TikTok may require different tips than users on Facebook – and promises to be an interesting field for future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jpubeco.2025.105345.

Data availability

Data will be made available on request.

References

Alesina, Alberto, Miano, Armando, Stantcheva, Stefanie, 2023. Immigration and redistribution. Rev. Econ. Stud. 90 (1), 1–39.

Ali, Shiza, Saeed, Mohammad Hammas, Aldreabi, Esraa, Blackburn, Jeremy, De Cristofaro, Emiliano, Zannettou, Savvas, Stringhini, Gianluca, 2021. Understanding the effect of deplatforming on social networks. In: Proceedings of the 13th ACM Web Science Conference 2021. pp. 187–195.

Allcott, Hunt, Gentzkow, Matthew, 2017. Social media and fake news in the 2016 election. J. Econ. Perspect. 31 (2), 211–236.

Allen, Jennifer, Arechar, Antonio A, Pennycook, Gordon, Rand, David G, 2021. Scaling up fact-checking using the wisdom of crowds. Sci. Adv. 7 (36), 1–10.

Allen, J., Watts, D.J., Rand, D.G., 2024. Quantifying the impact of misinformation and vaccine-skeptical content on facebook. Science 384 (6699), 1–8.

Bailey, Regan L, Gahche, Jaime J, Miller, Paige E, Thomas, Paul R, Dwyer, Johanna T, 2013. Why US adults use dietary supplements. JAMA Intern. Med. 173 (5), 355–361.

Bak-Coleman, Joseph B, Kennedy, Ian, Wack, Morgan, Beers, Andrew, Schafer, Joseph S, Spiro, Emma S, Starbird, Kate, West, Jevin D, 2022. Combining interventions to reduce the spread of viral misinformation. Nat. Hum. Behav. 6 (10), 1372–1380.

Barrera, Oscar, Guriev, Sergei, Henry, Emeric, Zhuravskaya, Ekaterina, 2020. Facts, alternative facts, and fact checking in times of post-truth politics. J. Public Econ. 182 (104123), 1–19.

Blair, Graeme, Imai, Kosuke, 2012. Statistical analysis of list experiments. Political Anal. 20 (1), 47–77.

Bode, Leticia, Vraga, Emily K., 2015. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. J. Commun. 65 (4), 619–638.

Brashier, Nadia M, Pennycook, Gordon, Berinsky, Adam J, Rand, David G, 2021. Timing matters when correcting fake news. Proc. Natl. Acad. Sci. 118 (5).

Broniatowski, David A, Simons, Joseph R, Gu, Jiayan, Jamison, Amelia M, Abroms, Lorien C, 2023. The efficacy of facebook's vaccine misinformation policies and architecture during the COVID-19 pandemic. Sci. Adv. 9 (37).

Bursztyn, Leonardo, Rao, Aakaash, Roth, Christopher, Yanagizawa-Drott, David, 2023. Opinions as facts. Rev. Econ. Stud. 90 (4), 1832–1864.

Carey, John M, Guess, Andrew M, Loewen, Peter J, Merkley, Eric, Nyhan, Brendan, Phillips, Joseph B, Reifler, Jason, 2022. The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, great britain and Canada. Nat. Hum. Behav. 6 (2), 236–243.

Cheung, Alan C.K., Slavin, Robert E., 2016. How methodological features affect effect sizes in education. Educ. Res. 45 (5), 283–292.

Chiou, Wen-Bin, Yang, Chao-Chin, Wan, Chin-Sheng, 2011. Ironic effects of dietary supplementation: illusory invulnerability created by taking dietary supplements licenses health-risk behaviors. Psychol. Sci. 22 (8), 1081–1086.

Clarke, Damian, Romano, Joseph P., Wolf, Michael, 2020. The romano–wolf multiple-hypothesis correction in stata. Stata J. 20 (4), 812–843.

Dai, Yue, Yu, Wenting, Shen, Fei, 2021. The effects of message order and debiasing information in misinformation correction. Int. J. Commun. 15, 1039–1059.

De Quidt, Jonathan, Haushofer, Johannes, Roth, Christopher, 2018. Measuring and bounding experimenter demand. Am. Econ. Rev. 108 (11), 3266–3302.

DellaVigna, Stefano, Gentzkow, Matthew, 2010. Persuasion: Empirical evidence. Annu. Rev. Econ. 2 (1), 643–669.

DellaVigna, Stefano, Kaplan, Ethan, 2007. The fox news effect: Media bias and voting. Q. J. Econ. 122 (3), 1187–1234.

Deslauriers, Louis, Schelew, Ellen, Wieman, Carl, 2011. Improved learning in a large-enrollment physics class. Science 332 (6031), 862–864.

Drexler, Alejandro, Fischer, Greg, Schoar, Antoinette, 2014. Keeping it simple: Financial literacy and rules of thumb. Am. Econ. Journal: Appl. Econ. 6 (2), 1–31.

Drolsbach, Chiara, Solovev, Kirill, Pröllochs, Nicolas, 2024. Community notes increase trust in fact-checking on social media. PNAS Nexus 3 (7), 217–230.

Ecker, Ullrich KH, Lewandowsky, Stephan, Cook, John, Schmid, Philipp, Fazio, Lisa K, Brashier, Nadia, Kendeou, Panayiota, Vraga, Emily K, Amazeen, Michelle A, 2022. The psychological drivers of misinformation belief and its resistance to correction. Nat. Rev. Psychol. 1 (1), 13–29.

Erbaugh, J.T., Chang, C.H., Masuda, Y.J., Ribot, J., 2024. Communication and deliberation for environmental governance. Annu. Rev. Environ. Resour. 49, 367–393.

Ershov, Daniel, Morales, Juan S., 2024. Sharing news left and right: Frictions and misinformation on Twitter. Econ. J. 134 (662), 2391–2417.

Fong, Jessica, Guo, Tong, Rao, Anita, 2024. Debunking misinformation about consumer products: Effects on beliefs and purchase behavior. J. Mark. Res. 61 (4), 659–681.

Freeman, Scott, Eddy, Sarah L, McDonough, Miles, Smith, Michelle K, Okoroafor, Nnadozie, Jordt, Hannah, Wenderoth, Mary Pat, 2014. Active learning increases student performance in science, engineering, and mathematics. Proc. Natl. Acad. Sci. 111 (23), 8410–8415.

Fryer Jr., Roland G., 2017. The production of human capital in developed countries: Evidence from 196 randomized field experiments. In: Handbook of Economic Field Experiments. vol. 2, Elsevier, pp. 95–322.

Gensing, Patrick, 2021. Wie die AfD angst vor impfungen schürt. Tagesschau. de.

Grinberg, Nir, Joseph, Kenneth, Friedland, Lisa, Swire-Thompson, Briony, Lazer, David, 2019. Fake news on Twitter during the 2016 US presidential election. Science 363 (6425), 374–378.

Gu, Jiayan, Dor, Avi, Li, Kun, Broniatowski, David A, Hatheway, Megan, Fritz, Lailah, Abroms, Lorien C, 2022. The impact of facebook's vaccine misinformation policy on user endorsements of vaccine content: An interrupted time series analysis. Vaccine 40 (14), 2209–2214.

Guess, Andrew M, Lerner, Michael, Lyons, Benjamin, Montgomery, Jacob M, Nyhan, Brendan, Reifler, Jason, Sircar, Neelanjan, 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. Proc. Natl. Acad. Sci. 117 (27), 15536–15545.

Guess, Andrew, Nagler, Jonathan, Tucker, Joshua, 2019. Less than you think: Prevalence and predictors of fake news dissemination on facebook. Sci. Adv. 5 (1).

Guess, Andrew, Nyhan, Brendan, Reifler, Jason, 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. Eur. Res. Counc. 9 (3), 1–14.

Gundersen, T., Alinejad, D., Branch, T.Y., Duffy, B., Hewlett, K., Holst, C., Owens, S., Panizza, F., Tellman, S.M., van Dijck, J., Baghramian, M., 2022. A new dark age? Truth, trust, and environmental science. Annu. Rev. Environ. Resour. 47 (1), 5–29.

Guriev, Sergei, Henry, Emeric, Marquis, Théo, Zhuravskaya, Ekaterina, 2023. Curtailing false news, amplifying truth. Work. Pap.

Henry, Emeric, Zhuravskaya, Ekaterina, Guriev, Sergei, 2022. Checking and sharing alt-facts. Am. Econ. Journal: Econ. Policy 14 (3), 55–86.

Hill, Carolyn J, Bloom, Howard S, Black, Alison Rebeck, Lipsey, Mark W, 2008. Empirical benchmarks for interpreting effect sizes in research. Child Dev. Perspect. 2 (3), 172–177.

Hopkins, Daniel J., 2009. No more wilder effect, never a whitman effect: When and why polls mislead about black and female candidates. J. Politics 71 (3), 769–781.

Jerit, Jennifer, Zhao, Yangzi, 2020. Political misinformation. Annu. Rev. Political Sci. 23, 77–94.

Kaiser, Tim, Lusardi, Annamaria, Menkhoff, Lukas, Urban, Carly, 2022. Financial education affects financial knowledge and downstream behaviors. J. Financ. Econ. 145 (2), 255–272.

Kaiser, Tim, Menkhoff, Lukas, 2022. Active learning improves financial education: Experimental evidence from uganda. J. Dev. Econ. 157.

Karlin, Beth, Zinger, Joanne F., Ford, Rebecca, 2015. The effects of feedback on energy conservation: A meta-analysis. Psychol. Bull. 141 (6).

Kuklinski, James H, Quirk, Paul J, Jerit, Jennifer, Schwieder, David, Rich, Robert F, 2000. Misinformation and the currency of democratic citizenship. J. Politics 62 (3), 790–816.

Lazer, David MJ, Baum, Matthew A, Benkler, Yochai, Berinsky, Adam J, Greenhill, Kelly M, Menczer, Filippo, Metzger, Miriam J, Nyhan, Brendan, Pennycook, Gordon, Rothschild, David, 2018. The science of fake news. Science 359 (6380), 1094–1096.

Lewandowsky, S., 2021. Climate change disinformation and how to combat it. Annu. Rev. Public. Health 42 (1), 1–21.

Lewandowsky, Stephan, Ecker, Ullrich K.H., Seifert, Colleen M., Schwarz, Norbert, Cook, John, 2012. Misinformation and its correction: Continued influence and successful debiasing. Psychol. Sci. the Public Interes. 13 (3), 106–131.

Lewandowsky, Stephan, Van Der Linden, Sander, 2021. Countering misinformation and fake news through inoculation and prebunking. Eur. Rev. Soc. Psychol. 32 (2), 348–384.

Liu, Xingyu, Qi, Li, Wang, Laurent, Metzger, Miriam J., 2023. Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. Commun. Res..

Loewenstein, George, Wojtowicz, Zachary, 2023. The economics of attention. Available At SSRN 4368304.

Luca, Michael, 2015. User-generated content and social media. In: Handbook of Media Economics. vol. 1, Elsevier, pp. 563–592.

Maaß, Sabrina, Wortelker, Jil, Rott, Armin, 2024. Evaluating the regulation of social media: An empirical study of the german netzdg and facebook. Telecommun. Policy 48 (5), 102719.

Maertens, Rakoen, Roozenbeek, Jon, Basol, Melisa, van der Linden, Sander, 2021. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. J. Exp. Psychology: Appl. 27 (1), 1.

Martel, Cameron, Rand, David G., 2024. Fact-checker warning labels are effective even for those who distrust fact-checkers. Nat. Hum. Behav. 8, 1957–1967.

Mitts, Tamar, Pisharody, Nilima, Shapiro, Jacob, 2022. Removal of anti-vaccine content impacts social media discourse. In: Proceedings of the 14th ACM Web Science Conference 2022. pp. 319–326.

Mummolo, Jonathan, Peterson, Erik, 2019. Demand effects in survey experiments: An empirical assessment. Am. Political Sci. Rev. 113 (2), 517–529.

Noar, Seth M., Benac, Christina N., Harris, Melissa S., 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions.. Psychol. Bull. 133 (4), 673.

Nyhan, Brendan, 2021. Why the backfire effect does not explain the durability of political misperceptions. Proc. Natl. Acad. Sci. 118 (15), e1912440117.

Nyhan, Brendan, Porter, Ethan, Reifler, Jason, Wood, Thomas J, 2020. Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. Political Behav. 42 (3), 939–960.

Nyhan, Brendan, Reifler, Jason, 2010. When corrections fail: The persistence of political misperceptions. Political Behav. 32 (2), 303–330.

Nyhan, Brendan, Reifler, Jason, 2015. Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. Vaccine 33 (3), 459–464.

Pennycook, Gordon, Epstein, Ziv, Mosleh, Mohsen, Arechar, Antonio A, Eckles, Dean, Rand, David G, 2021. Shifting attention to accuracy can reduce misinformation online. Nature 592 (7855), 590–595.

Pennycook, Gordon, McPhetres, Jonathon, Zhang, Yunhao, Lu, Jackson G, Rand, David G, 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychol. Sci. 31 (7), 770–780.

Pennycook, Gordon, Rand, David 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, Gordon, Rand, David G., 2021. The psychology of fake news. Trends Cogn. Sci. 25 (5), 388–402.

Radimer, Kathy, Bindewald, Bernadette, Hughes, Jeffery, Ervin, Bethene, Swanson, Christine, Picciano, Mary Frances, 2004. Dietary supplement use by US adults: data from the national health and nutrition examination survey, 1999–2000. Am. J. Epidemiol. 160 (4), 339–349.

Romano, Joseph P., Wolf, Michael, 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.

Rooney, Brenda L., Murray, David M., 1996. A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. Health Educ. Q. 23 (1), 48–64.

Roozenbeek, Jon, van der Linden, Sander, Goldberg, Beth, Rathje, Steve, Lewandowsky, Stephan, 2022. Psychological inoculation improves resilience against misinformation on social media. Sci. Adv. 8 (34), eabo6254.

Ruiz-Primo, Maria Araceli, Briggs, Derek, Iverson, Heidi, Talbot, Robert, Shepard, Lorrie A, 2011. Impact of undergraduate science course innovations on learning. Science 331 (6022), 1269–1270.

Stantcheva, Stefanie, 2021. Understanding tax policy: How do people reason? Q. J. Econ. 136 (4), 2309–2369.

Swire, Briony, Berinsky, Adam J, Lewandowsky, Stephan, Ecker, Ullrich KH, 2017. Processing political misinformation: comprehending the trump phenomenon. R. Soc. Open Sci. 4 (3).

Swire-Thompson, Briony, Cook, John, Butler, Lucy H, Sanderson, Jasmyne A, Lewandowsky, Stephan, Ecker, Ullrich KH, 2021. Correction format has a limited role when debunking misinformation. Cogn. Research: Princ. Implic. 6 (83), 1–15.

Swire-Thompson, B., Lazer, D., 2020. Public health and online misinformation: challenges and recommendations. Annu. Rev. Public. Health 41 (1), 433–451.

Tversky, Amos, Kahneman, Daniel, 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. Science 185 (4157), 1124–1131.

Verplanken, Bas, Orbell, Sheina, 2022. Attitudes, habits, and behavior change. Annu. Rev. Psychol. 73, 327–352.

Vosoughi, Soroush, Roy, Deb, Aral, Sinan, 2018. The spread of true and false news online. Science 359 (6380), 1146–1151.

Vraga, Emily K., Bode, Leticia, 2017. Using expert sources to correct health misinformation in social media. Sci. Commun. 39 (5), 621–645.

Zhuravskaya, Ekaterina, Petrova, Maria, Enikolopov, Ruben, 2020. Political effects of the internet and social media. Annu. Rev. Econ. 12, 415–438.