News credibility labels have limited but uneven effects on news diet quality and fail to reduce misperceptions*

Kevin Aslett[a], Andrew M. Guess[b], Jonathan Nagler[a,c], Richard Bonneau[a,d], and Joshua A. Tucker[a,c]

[a]Center for Social Media and Politics, New York University
[b]Department of Politics, Princeton University
[c]Wilf Family Department of Politics, New York University
[d]Department of Biology, New York University

Abstract

As the primary arena for viral misinformation shifts toward transnational threats such as the Covid-19 pandemic, the search continues for scalable, lasting countermeasures compatible with principles of transparency and free expression. To advance scientific understanding and inform future interventions, we conducted a randomized field experiment evaluating the impact of source credibility labels embedded in users' social feeds and search results pages. By combining representative surveys ($N = 3,337$) and digital trace data ($N = 946$) from a subset of respondents, we provide a rare ecologically valid test of such an intervention on both attitudes and behavior. On average across the sample, we are unable to detect changes in real-world consumption of news from low-quality sources after three weeks, and we can rule out even small effects on perceived accuracy of popular misinformation spread about the Black Lives Matter movement and Covid-19. However, we present suggestive evidence of a substantively meaningful increase in news diet quality among the heaviest consumers of misinformation in our sample. We discuss the implications of our findings for our understanding of the determinants of news diets and for practical questions about designing interventions to counteract online misinformation.

*We are extremely grateful to Craig Newmark Philanthropies for supporting this research project. The Center for Social Media and Politics at New York University is generously supported by funding from the National Science Foundation, the John S. and James L. Knight Foundation, the Charles Koch Foundation, the Hewlett Foundation, Craig Newmark Philanthropies, the Siegel Family Endowment, and NYU's Office of the Provost and Global Institute for Advanced Study. NewsGuard did not consult with the authors on the study design or provide funding support for this research. This study has been approved by the Princeton University Institutional Review Board (#12800) and the New York University Institutional Review Board (#FY2020-4278). Special thanks to Sam Luks at YouGov for facilitating data collection. We thank Tali Mendelberg for helpful feedback.

The internet and social media have drastically decreased the cost of disseminating information by reducing reliance on traditional gatekeepers. As a consequence of this openness and availability, news and information sources have flourished from a variety of ideological and cultural perspectives. The resulting cacophony has encouraged participation by previously underrepresented voices and enabled criticism of dominant authorities. At the same time, it has intersected with existing political divides in ways that have contributed to pathologies in American political discourse including the spread of misinformation (Lazer et al. 2018; Vosoughi et al. 2018; Grinberg et al. 2019; Guess et al. 2019; Osmundsen et al. ming), disagreements about basic facts related to governance and policy (Flynn et al. 2017; Anspach et al. 2019; Pennycook and Rand 2021), and lowered trust in established media (Guess et al. 2021). Of particular concern is the possibility that these problems are interlinked: As political divisions widen, partisan media alienate people from authoritative sources, which could make it more difficult to counteract potentially corrosive — and in the case of public health during a pandemic, life-threatening (Brennen et al. 2020) — misinformation.

Over the past several years, scholars, technologists and policy makers have proposed a number of solutions intended to reduce exposure to misleading information. These range from relatively intrusive measures such as algorithmic downranking, to subtle warnings and labels targeted at specific factual claims (Ecker et al. 2010; Clayton et al. 2019), to general efforts to boost digital media literacy skills or accuracy motivations (Guess et al. 2020b; Pennycook et al. 2021). A key challenge in these efforts is how to balance the strength of an intervention with potential negative externalities in the form of unintended spillover effects (Pennycook et al. 2020; Nyhan et al. 2013) or limits on individual autonomy and freedom of expression. With this tension in mind, we focus on simple feedback in the form of informational labels designed to educate people about the quality of sources that they consume and view in their search or social media feeds (Lorenz-Spreen et al. 2020). This approach builds on humans' tendency to rely on cognitive shortcuts and heuristics, which depending on context can be relatively informative (in the case of source transparency; see Iyengar and Hahn 2009; Gigerenzer and Selten 2002; Pennycook and Rand 2019) or potentially distorting (in the case of social cues, which are a common feature of social media; see Messing and Westwood 2014). It also shifts attention to accuracy, a strategy shown to increase the average quality of news that individuals share on social media (Pennycook et al. 2021).

Aside from providing critical policy-relevant evidence, the efficacy of this type of dynamic feedback sheds light on the determinants of people's information diets. One set of arguments contends that people's online news consumption is largely determined by their political preferences — a perspective that, when translated to the domain of politics, implies that many people reside in relatively impermeable, non-overlapping informational bubbles (Sunstein 2017). In this view, political judgments as well as factual beliefs could be powerfully shaped by strong and persistent differences in partisans' news diets. A revisionist literature has

emerged challenging many of these claims, which suggests that most people (with the exception of some strong partisans) regularly encounter cross-cutting information in their online browsing activity and social media feeds, and consumption habits are less ingrained than they first appeared. (Gentzkow and Shapiro 2011; Bakshy et al. 2015; Eady et al. 2019; Fletcher et al. 2020; Guess 2021). If news consumers are receptive to informational nudges, this suggests that media consumption habits are driven at least in part by environmental factors, such as online choice architecture, that can be altered (Thaler and Sunstein 2009).

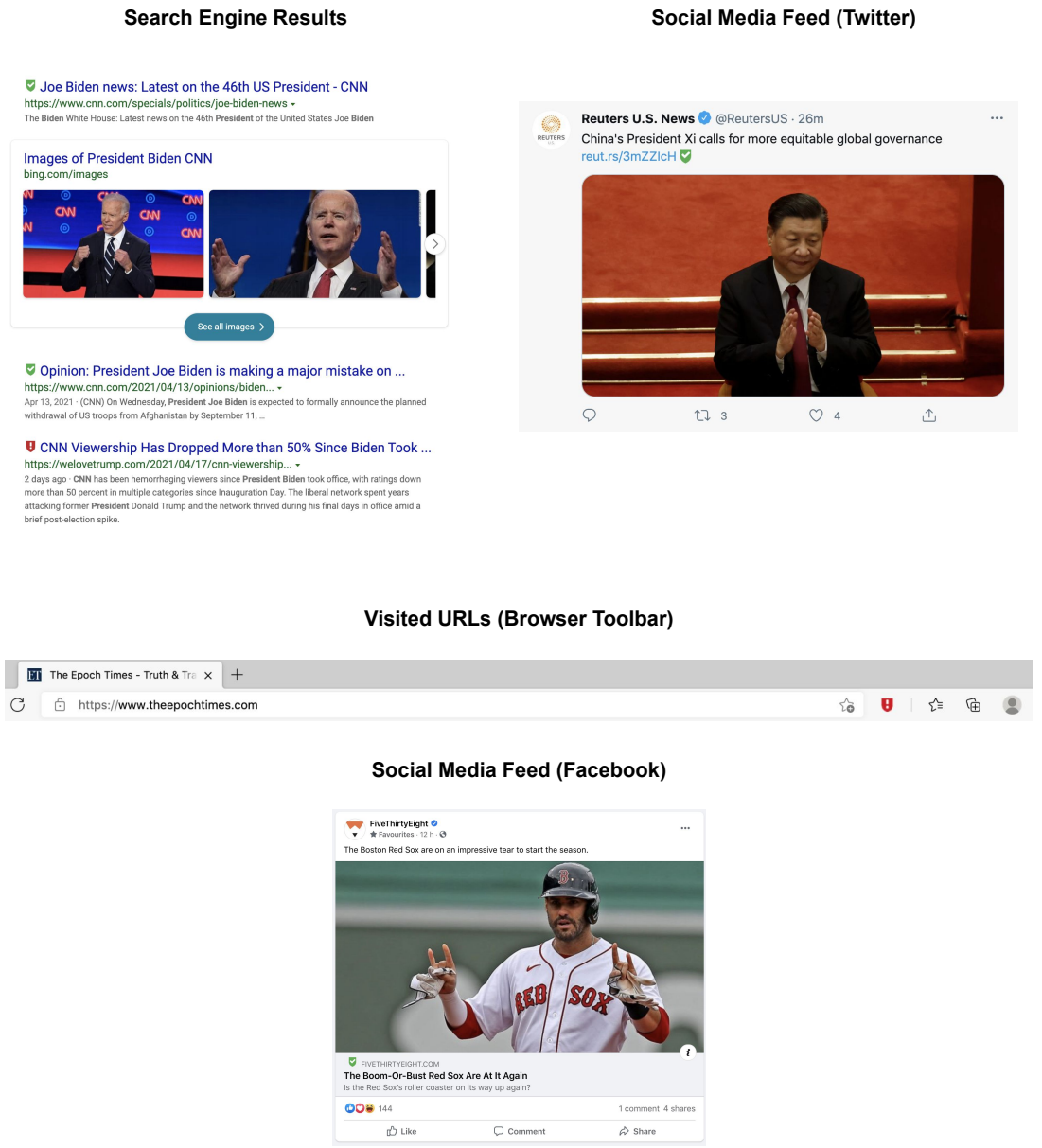
Prior research suggests mixed expectations about whether subtle, embedded source information cues can change online news consumption behavior or reduce misperceptions. As suggested, there is limited direct evidence that relatively light-touch interventions designed to pre-bunk or directly challenge misinformation can affect downstream online behavior (though for an important exception related to link sharing on Twitter, see Pennycook et al. 2021). At the same time, behavioral-science approaches suggest potential benefits of quality heuristics in limited-information environments (Lupia 1994; Gigerenzer et al. 2011; Lorenz-Spreen et al. 2020). Looking beyond potential news diet changes, source-level ratings could possibly reduce belief in misleading claims given that prior research has found that source-level information from experts can reduce belief in specific articles from news sources rated as unreliable (Kim et al. 2019; Kim and Dennis 2019). But even if perceived accuracy of false claims is reduced as a result of our intervention, this effect might be accompanied by a corresponding decrease in perceived accuracy of true claims as well, as studies have sometimes found (e.g., Guess et al. 2020b).

To study this approach, we build on recent innovations for rigorously evaluating online tools (Munzert et al. 2021a,b). Within a two-wave survey panel, we randomly encouraged participants to install a prominent web browser extension, NewsGuard, which embeds straightforward source-level indicators of news reliability into users' search engine results pages (SERPs), social feeds and visited URLs.[1] Different "shield" symbols are placed in-feed to provide visual summaries of sources' quality. A green shield indicates a reliable source (examples include CNN, Fox News, and *The Washington Post*), a red shield indicates an unreliable source (examples include Gateway Pundit, *Epoch News*, and Daily Kos), a gray shield indicates a source with user-generated content (such as YouTube, Wikipedia, and Reddit), and a gold shield represents satire (such as *The Onion*, *Babylon Bee*, and *The Daily Mash*). The user can click on the shield to see an overlay of more detailed information about the reliability of the news domain in question.[2]

[1] NewsGuard launched in 2018 and produces ratings based on neutral criteria evaluated by a team of journalists and editors; more information can be found in the Materials and Methods section and at `www.newsguardtech.com`. Examples of how the source ratings appear to news consumers in different settings are shown in Fig. 1. Although this study employed the NewsGuard extension, which was freely available in app stores at the time of fielding, NewsGuard did not provide any financial support or assistance in the design of this study.

[2] We show these source reliability symbols in Section G of the Supplementary Materials.

Figure 1: NewsGuard source labels.



We ran a pre-registered[3] field experiment, drawing on a representative online sample of Americans. Our main hypotheses test whether in-feed source reliability labels shift downstream news and information consumption from unreliable sources known for publishing misleading or false content to more reliable sources (H1), increase trust in mainstream media[4] and reliable sources (H2), and mitigate phenomena associated with

[3]The pre-registration can be found here: `https://osf.io/9qrkt/`

[4]We also test if the effect of the treatment on trust in media is larger among individuals who have lower levels of initial trust in media.

democratic dysfunction (affective polarization and political cynicism)[5] (H3). We also consider three research questions for which our *a priori* expectations were less clear. First, past research suggests that certain kinds of interventions can reduce people's beliefs in both accurate and inaccurate information (Clayton et al. 2019; Guess et al. 2020b), so we examine whether respondents encouraged to install the NewsGuard extension were more or less likely to believe popular false and true stories that spread during the treatment period.[6] Second, we explore whether downstream effects are observable on other outcomes such as trust in institutions, belief that "fake news" is a problem in general, and belief that "fake news" is a problem in the mainstream media. Third, we explore whether any of the identified effects are greater among subgroups found in prior research to more frequently engage with online misinformation.[7] Results from all of our pre-registered analyses can be found in the Supplementary Materials, Sections C, D, E, and F, G and H.

Combining panel survey data and individual-level web visit data, we find that in-browser contextual source labels: (1) do not measurably shift participants' online consumption from unreliable sources known for publishing misleading or false content to more reliable sources; (2) fail to reduce average belief in widely circulated inaccurate claims; and (3) do not alter trust in the media generally. Our estimates — especially for survey-based outcomes — are well-powered, and we can rule out even very small Intent-to-Treat effects (Cohen's $d < 0.07$). Our null findings on changes in participants' information diet quality are somewhat noisier, though we can still rule out small effects (Cohen's $d < 0.09$) according to standard benchmarks. Our noisier estimates on behavior are themselves informative, since they reflect the well-established reality that people with diets consisting overwhelmingly of untrustworthy news sources are a relatively small subset of the population (Guess et al. 2020a). Interventions designed to improve news quality through dynamic feedback may therefore need to focus their efforts on these individuals and tailor the information they provide accordingly. Consistent with this – and in light of our original null findings – we undertake supplementary analyses (not preregistered) and find that the treatment actually does appear to improve the average reliability score of the news consumed by participants at the lowest decile of pre-treatment online news diet quality.

Results

To measure the effect of these source labels, we fielded a two-wave panel survey in summer 2020 (Wave 1: May 28–June 9, $N = 3,862$; Wave 2: June 19–June 30, $N = 3,337$) that included a randomized incentive to

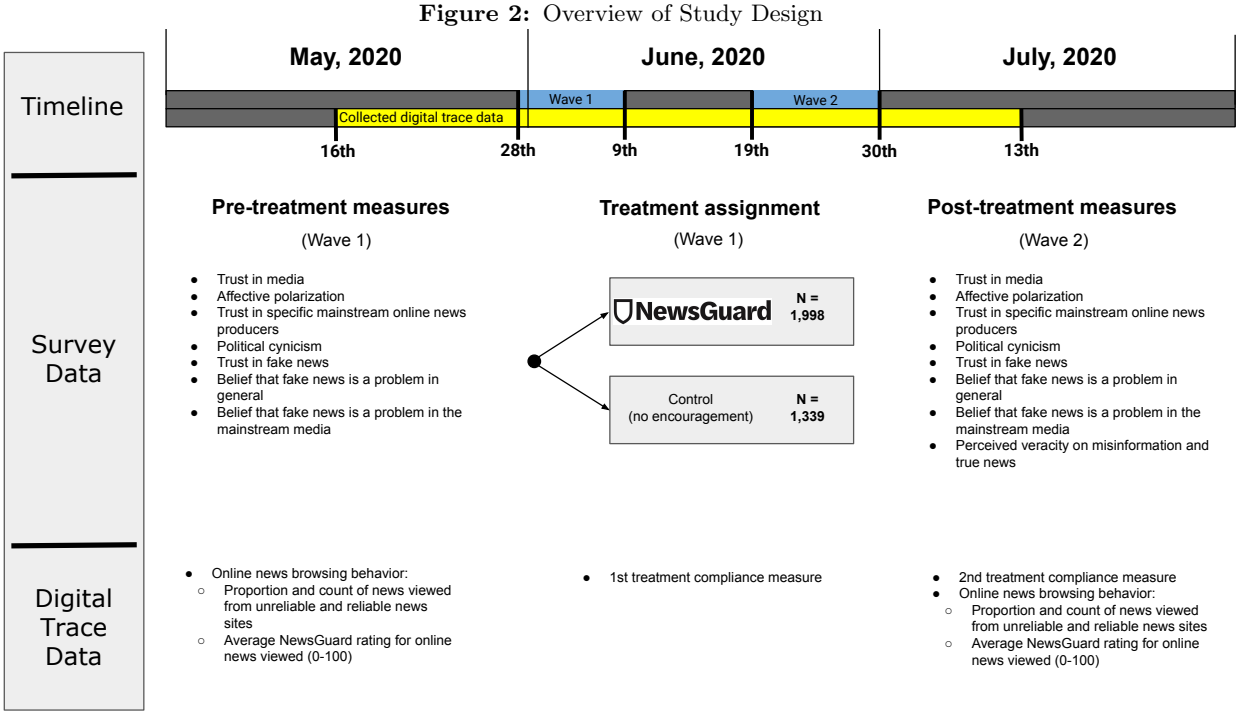
[5] We also test if the effect of the treatment on affective polarization is larger among individuals who have higher levels of affective polarization.

[6] We were not able to pre-register this research question because we selected the items as close to fielding as possible. We include the results because they are more directly comparable to studies evaluating the effects of interventions designed specifically to reduce misperceptions.

[7] These groups include those who use social media sites more frequently, have low levels of digital literacy, consume more news, and already visit more online publishers of untrustworthy news.

install the NewsGuard web extension at the beginning of the first wave. Enough respondents were recruited that we are confident that we could detect even a small standardized effect size (9% of a standard deviation) among respondents for whom we have behavioral data as well as the larger survey sample. Results from our power analyses can be found in the Supplementary Materials in Section J. Fig. 2 presents an overview of the study design. In addition to studying survey-based outcomes, we analyze linked digital trace data to measure the quality of news consumption of a subset of our participants. We create five distinct measures of news diet quality[8] over three time periods: (i) the period before a respondent was assigned treatment in the Wave 1 survey; (ii) the 3- to 4-week period from treatment assignment (May 28–June 9) to June 30; (iii) the nearly two-week period from July 1 to July 13. Testing the effect of this treatment on news consumption during the third period was not a part of our original set of pre-registered hypotheses, but rather a pre-registered research question. Using the measurement of news consumption in the the third period we leverage the exogenous disabling of NewsGuard's free capabilities on July 1[9] to determine whether the behavioral effects of this intervention decay after its features are no longer available (Gerber et al. 2011), or if the intervention has more durable effects like other novel informational nudges (Coppock et al. 2018).

Figure 2: Overview of Study Design



[8]We calculated the average NewsGuard reliability score for websites visited, proportion and counts of unreliable (NewsGuard score < 60) news sites visited, and proportion and counts of reliable news sites visited.

[9]On that day, the NewsGuard extension became a pay service with a monthly subscription fee of $2.99. At that point, those who did not sign up to purchase the extension continued to encounter shield icons next to news stories, but the color no longer reflected credibility and contextual information was no longer accessible. Given the disabling of NewsGuard's free capabilities on July 1, respondents were effectively treated for 3 to 4 weeks.

In this section we primarily report covariate-adjusted estimates of an Intent-to-Treat (ITT) effect measuring differences between the control and treatment groups on outcomes of interest. In the Supplementary Materials (Section C) we report covariate-adjusted estimates of Complier Average Causal Effects (CACE), the effect among those who installed the browser extension as a result of the randomized treatment. To verify treatment compliance, we developed an automated script linked in the survey that measured whether participants in the treatment and control groups had installed and activated the NewsGuard extension on their web browsers twice: directly after the treatment was assigned in Wave 1 and in the last week of the treatment period. 95% of respondents in the treatment group passed the first compliance check and 80% passed both the first and second compliance check. For most demographic characteristics we find no statistically significant evidence that respondents who comply are very much different than those who did not comply; moreover, for demographic characteristics where we find a statistically significant difference, the magnitudes of these differences are very small.[10]

Since a lack of statistically significant coefficients does not necessarily imply effects of a negligible magnitude, we use equivalence testing to rule out any meaningful effects in all of the following models (Rainey 2014; Lakens 2017; Hartman and Hidalgo 2018). Conservatively, negligible effects are defined as those smaller than 20% of a standard deviation (Cohen 1969) of the population on a pre-treatment measure of a variable, although others have advocated for higher thresholds (Imbens and Rubin 2015; Hartman and Hidalgo 2018). We thus calculate standardized effects for each estimate shown in Fig. 3 and 4. The magnitudes of these standardized effects never rise above 10% of a standard deviation, thus remaining far below the most conservative established threshold, rejecting the hypothesis of even a small effect on the outcome variable. In addition to these standardized effect sizes, we also estimate minimum detectable effects for covariate-adjusted ITT models reported in the main text in Section J in the Supplementary Materials. Assuming power of 0.80 and that statistically significant effects must cross the p-value threshold of 0.05, the models can at a minimum detect effects larger than 6.5% of a standard deviation for all of the attitudinal measures and any effect larger than 8.9% for all of the behavioral measures. Our design is thus well-powered to detect small effects of the intervention.

In our sample of respondents for whom we collect behavioral news consumption data ($N = 946$), most did not visit an unreliable news site during the two- to three-week pre-treatment period (over 65%) and just under 12% of our sample's news diet consisted of at least 5% of visits to news sites deemed unreliable by NewsGuard. Indeed, the average NewsGuard reliability score of respondents' news diets was 87.6 out of 100, and only 1.5% of respondents' news diets had an average NewsGuard reliability score below 60 (the threshold

[10]More details on how compliance was estimated, compliance rates, and how compliers differed from non-compliers can be found in the Materials and Methods section. About 1% of those in the control group already had the NewsGuard extension installed and activated.

for reliability). This long tail of low-quality online news consumption is in line with what previous studies have reported (Guess et al. 2020a; Paskhalis et al. 2021), though the relatively rare prevalence of visits to unreliable sites among our respondents creates a challenge in terms of statistical power when estimating treatment effects on this behavior. For this reason, we also test the effect of the treatment on the subset of respondents who consume the most low-quality news to determine if this intervention is effective among those who consume higher levels of low-quality news.

Contrary to Hypothesis 1, we do not find that randomized exposure to in-browser source reliability information shifts online consumption of news away from unreliable publishers. The treatment effect estimates for each pre-registered behavioral outcome are presented in Fig. 3. As the figure indicates, we do not find statistically significant decreases in the proportion of news consumed from unreliable sources or in the count of unreliable online news consumed, either during the period when NewsGuard was installed by those in the treatment group or in the two weeks after NewsGuard became a pay service. We also do not find that the intervention measurably shifts individuals toward reliable news: We do not observe statistically significant increases in the proportion of news consumed from reliable sources, in the count of visits to these sources, or in the average reliability score of online news consumed in either time period due to the treatment. In addition to an absence of statistically significant effects, the estimated magnitudes are extremely small: all of the reported effect sizes constitute less than a 0.08 change in the standard deviation of the pre-treatment measure of that variable.

Given these results we conducted a series of non-preregistered supplementary analyses to better understand the seemingly negligible impact of the news quality labels on news consumption. To determine if the treatment affects the time spent on unreliable or reliable news websites, we weight each variable by duration, but this does not measurably change the results (details can be found in Section C of the Supplementary Materials). We also measured these variables strictly as referrals from social media sites and search engines and find no statistically significant treatment effects (details can be found in Section C of the Supplementary Materials). Finally, we divide the sample according to pre-treatment deciles of respondents' average news reliability scores. Interestingly, we find relatively strong and statistically significant treatment effects among those in the lowest decile of news diet reliability (Fig. 4): Relative to the average pre-treatment value, we estimate a 5.4% increase in the treatment period and a 8.6% increase beginning July 1 in the average reliability score of news consumed (full results for each behavioral measure across each decile can be found in Section K of the Supplementary Materials). Though suggestive, these powerful subgroup effects are consistent with the intervention being mainly effective among those who consume the greatest amount of content from relatively unreliable sources.

Figure 3: This figure presents estimates of the effect of the intervention (CACE, with 95% confidence intervals) on our pre-registered online behavioral measures in the two periods after treatment assignment: before July 1 when the NewsGuard extension was freely available and the two-week period between July 1–13 when the NewsGuard extension was disabled. The effect is reported in standard deviations of that measure (pre-treatment).

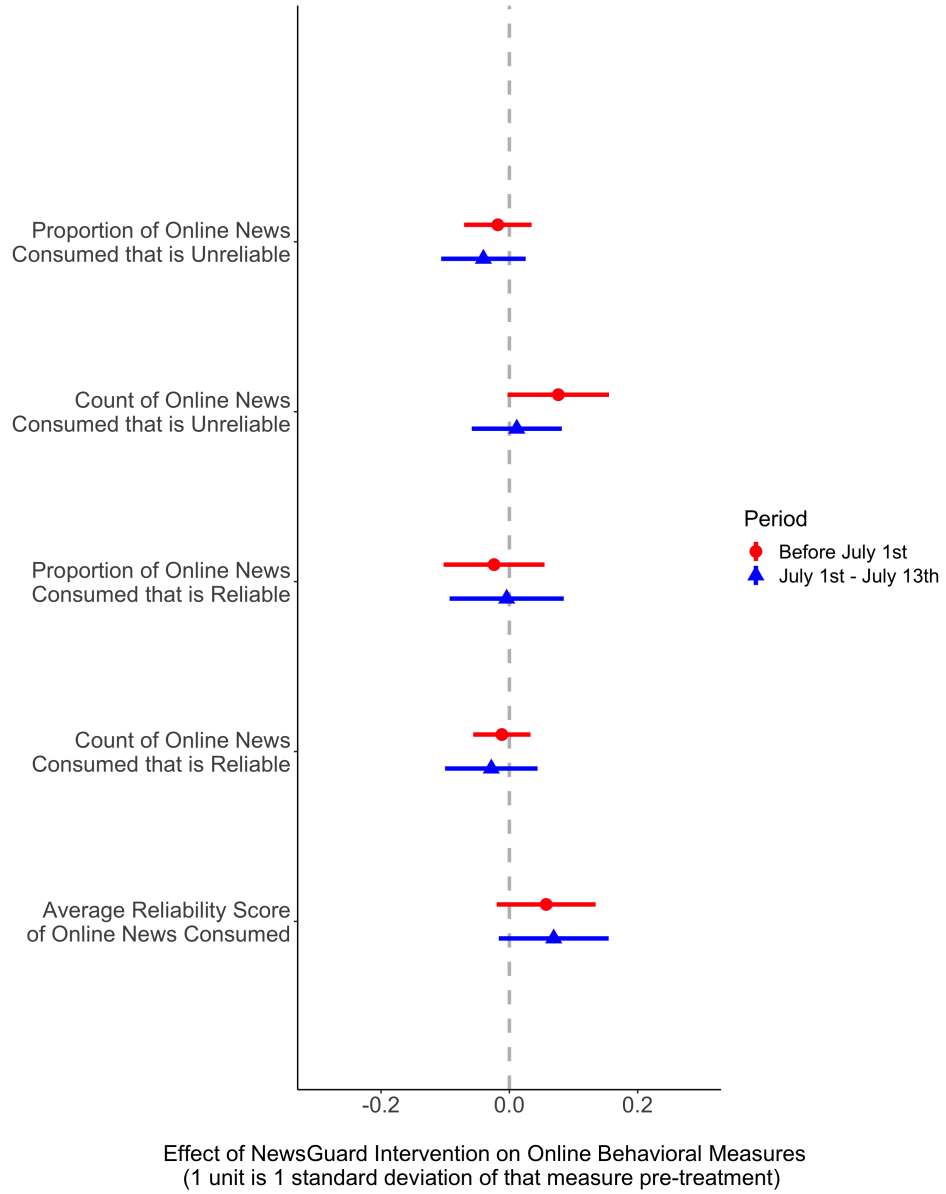
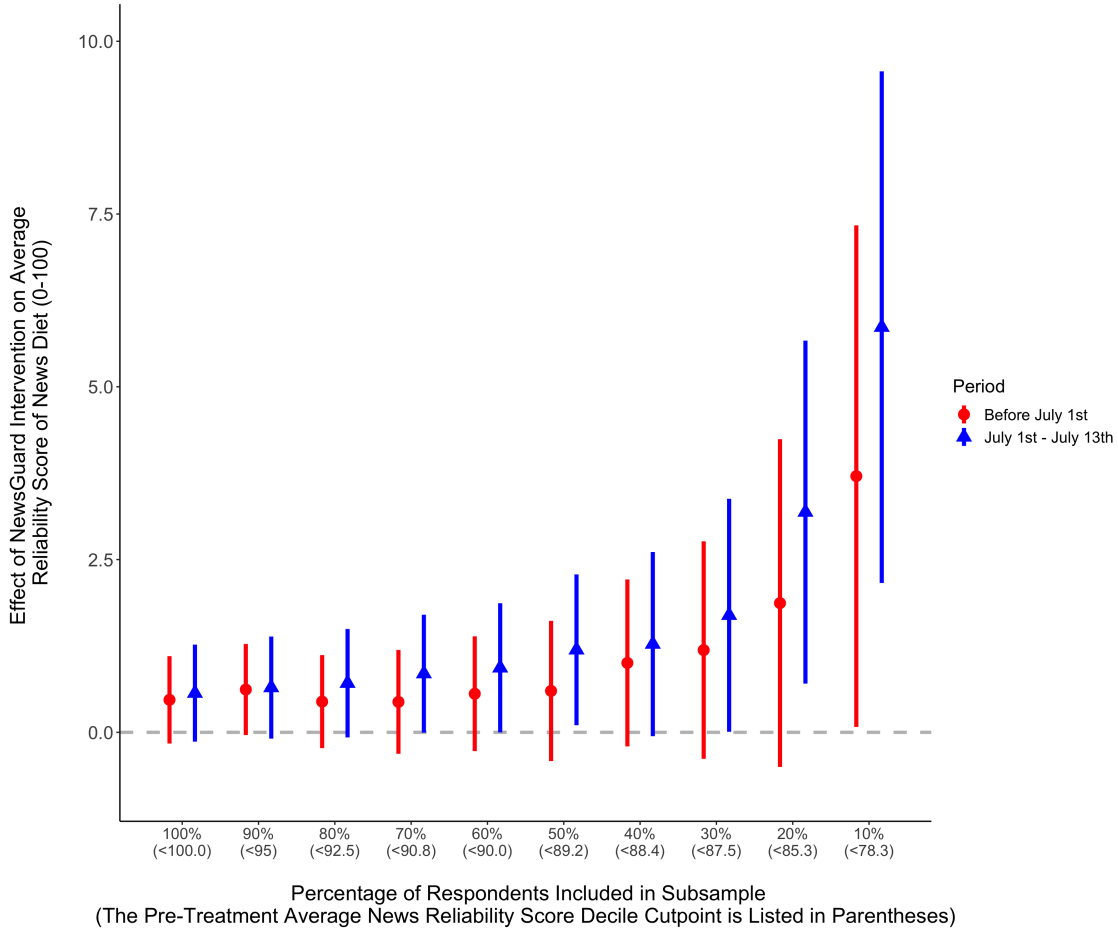


Figure 4: This figure presents estimates of the effect of the intervention on the average NewsGuard reliability score of respondents' news diets (with 95% confidence intervals). This is estimated for subsets of the sample with pre-treatment reliability scores below successive decile cutpoints.



The next hypothesis predicted that source reliability feedback would increase trust in the media and reliable sources (H2). We find that the treatment does not increase trust in media or in specific reliable news sources. Our estimates of the treatment effect on trust in media and in selected specific sources deemed reliable by NewsGuard are not statistically distinguishable from zero, and the reported effect sizes are less than a 0.05 change in the standard deviation of the pre-treatment measure of that variable. A figure depicting the effect of the intervention on our pre-registered attitudinal measures can be found in Section K of the Supplementary Materials.

We also find no support for our final hypothesis, which predicted that exposure to the treatment could help to alleviate pathologies such as affective polarization[11] and political cynicism associated with consuming, believing, and sharing news from unreliable sources (H3), which in retrospect is perhaps not surprising because we did not find an effect for the treatment on consuming of believing news from unreliable sources.

[11] We also do not find that any effect of the treatment is concentrated among those with higher levels of affective polarization. These results can be found in Sections E and F in the Supplementary Materials.

RQ1 asked whether source reliability information affects belief in misinformation as well as accurate claims. To answer this question, all respondents were asked to judge the veracity of five widely circulated statements about the Black Lives Matter (BLM) movement and five similarly well-circulated statements about Covid-19 using a four-point scale in Wave 2. Of the five statements about each topic, three were false and two were true. By taking the mean of the perceived veracity measure for the three false statements about each topic we created a measure of belief in misinformation in the BLM movement and Covid-19, and we likewise created a measure of belief in true information using the other two items. The intervention *had no effect on* belief in misinformation about the Black Lives Matter movement and Covid-19, and it did not measurably affect belief in the true statements.[12]

Our second research question asked whether exposure to the intervention leads to effects on other outcomes such as trust in institutions, belief that fake news is a problem in general, and belief that fake news is a problem in the mainstream media. We do not find that the intervention measurably affected these outcomes.[13] Finally, RQ3 asked if any effects are moderated by specific characteristics (proportion of news consumed that is unreliable, partisanship of news diet, online news consumption, social media use, digital literacy, and belief in scientific misinformation), but we find no other consistent evidence of effect heterogeneity.

Discussion

Despite the promise of browser-based tools designed to reduce users' reliance on misinformation,[14] evidence from a pre-registered randomized field experiment among a large representative sample of Americans reveals that the particular intervention studied here — providing dynamic, in-feed source reliability labels — does not measurably improve news diet quality or reduce misperceptions on average among the general population. Our estimates, based on both survey and behavioral data collected over an extended period, are precise and rule out even modest effect sizes by conventional standards.

Though we are able to measure pre- and post-treatment news consumption "in the wild" as well as treatment compliance at the point of installation, we do not capture the contents of participants' browser experience — what they *see* but don't click on. This means that we did not directly observe how often users encountered NewsGuard's labels in the course of browsing their feeds and perusing search results pages during the study period. We can estimate how often they were exposed to these labels by counting the number of visits to online news sites rated by NewsGuard (which trigger a source quality label in the browser bar when

[12] Reported effect sizes are all below a 0.03 change in the standard deviation.
[13] Reported effect sizes are all below a 0.03 change in the standard deviation.
[14] See: `https://www.politico.com/news/magazine/2021/04/27/america-social-media-problem-newsguard-484757`

the extension is installed), Google search results pages, and time spent on Facebook or Twitter during the pre-treatment period; we present distributions of these measures in the Supplementary Materials in Section J. We are most interested in respondents' visits to search engine results pages (SERPs) and time spent on social media feeds because users scrolling through these pages are likelier to be exposed to the credibility ratings for multiple news sites as opposed to one at a time. This should theoretically translate to greater potential for the labels' effectiveness than discrete visits to online news sites rated by NewsGuard, though we find that the majority of our respondents rarely visit Google SERPs and spend relatively limited time on Facebook or Twitter.[15] The relative rarity of exposure to credibility ratings in these contexts may partly explain the negligible effect of this intervention on the quality of online news consumption and downstream attitudes.

Our results speak to a large body of work on heuristics and cognitive processing (e.g., Chaiken 1987; Flanagin and Metzger 2000). The NewsGuard shield icons provide expertise cues, which have been shown since the work of Hovland to be associated with increased source credibility (Hovland et al. 1953; Kim et al. 2019; Kim and Dennis 2019). However, in a more partisan age in which attitudes toward news sources are strongly correlated with partisanship (Jurkowitz et al. 2020), relatively subtle contextual information may not be a sufficiently powerful prod to shift perceptions of source credibility. Our findings also relate to an emerging body of research on accuracy motives among social media users and attempts to encourage discernment in news sharing behavior via "accuracy nudges" (Pennycook et al. 2021). Though real-world experimentation points to the efficacy of such interventions, our null findings raise the possibility that accuracy mechanisms may operate differently for publicly observable sharing behavior than for private consumption. For example, sharing false information may come with a perceived reputational cost that is not associated with low-quality news consumption (Altay et al. 2019).

Although we do not uncover statistically significant average treatment effects, we present suggestive evidence of a substantively meaningful boost in news quality among the heaviest consumers of misinformation in our sample (who comprise a small proportion of respondents, approximately 10%, consistent with prior research on fake news exposure). In both the treatment period (before July 1) and in the post-treatment period (beginning July 1) we observed the quality of news diets among the lowest 10% of our sample increase by 5.4% and 8.6% from their pre-treatment levels, respectively. For the period beginning July 1, we even observe a statistically significant increase in news reliability among those in the bottom 20% of pre-treatment news consumption quality.

This finding illustrates the challenge of studying interventions aimed at specific (and often difficult to

[15]For example, while 92% of respondents visited Facebook at least once during the pre-treatment period, 58% spent more than an hour on the platform and 26% spent over 5 hours in total during that period.

sample) populations. From a policy perspective this intervention is arguably designed to change the behavior of a specific population, the heaviest consumers of misinformation, so the absence of an effect overall on the general population does not necessarily indicate that the tool is ineffective. Hypothetically, generalizing our estimates of treatment effects among the heaviest consumers of misinformation could imply downstream consequences for publishers as well, as the demand for news and information from unreliable outlets might decrease (an effect which could, in turn, be amplified through search ranking and social media recommendation engines). A sizable drop in traffic, and the resulting revenue loss, could remove some of the financial incentives for producing misinformation as well as deter future entrants into the market. Our estimates for those who consume the most misinformation suggest that future studies should explore oversampling from the populations whose behavior an intervention is designed to change. Our results also suggest that those who develop and promote such interventions (including browsers and platforms themselves) should more precisely target these populations.

Materials and Methods

NewsGuard Extension and Ratings

To produce credibility ratings, NewsGuard employs a team of trained journalists and editors to review and rate news and information websites based on nine journalistic criteria. The criteria assess basic practices of reliability and transparency. Based on a site's performance on these nine criteria, it is assigned a reliability rating from 0–100.[16] Online domains with score of 60 or higher are considered reliable (green shield), while scores below 60 are considered unreliable.[17] A histogram of NewsGuard scores for the majority of online news domains can be found in Section B of the Supplementary Materials.[18] NewsGuard can be installed on all major web browsers (Safari, Microsoft Edge, Mozilla Firefox, Internet Explorer, and Google Chrome) as well as Android and iOS mobile phones. Normally, the NewsGuard extension costs $2.99 per month, but it is available for free (and bundled) with Microsoft Edge as well as to over 200 million potential users worldwide through assorted partnerships.[19]

[16]The criteria are: "Does not repeatedly publish false content", "Gathers and presents information responsibly", "Regularly corrects or clarifies errors", "Handles the difference between news and opinion responsibly", "Avoids deceptive headlines", "Website discloses ownership and financing", "Clearly labels advertising", "Reveals who's in charge, including any possible conflicts of interest", and "The site provides names of content creators, along with either contact or biographical information".

[17]Over 41% of the more than 5,000 news domains rated received a red, suspect rating. Reassuringly, the NewsGuard list contains most of the "fake news" publishers identified by Allcott et al. (2019); 88% of the online news domains in the Allcott et al. (2019) list are rated as unreliable by NewsGuard. In addition, 99% of mainstream online news domains identified by Microsoft Project Ratio are rated as reliable by NewsGuard.

[18]Over the course of 2020, NewsGuard rated 2,144 additional news domains and partnered with the World Health Organization to report misinformation and flag 371 websites that spread misinformation about Covid-19 in the first months of the pandemic.

[19]Currently, NewsGuard is offered for free to 30 million BT internet and mobile customers, students through TurnItIn, and patrons at over 750 libraries (including the Chicago Public Library).

Data, Sample and Measures

We conducted a panel survey of U.S. adults that included an encouragement to install NewsGuard in the first wave. This two-wave online panel survey was fielded by the survey company YouGov in the summer of 2020 (Wave 1: May 28–June 9, $N = 3,862$; Wave 2: June 19–July 1, $N = 3,337$). Respondents were selected by YouGov's matching and weighting algorithm to approximate the demographic and political attributes of the U.S. population (32% college graduates, 45% male, median age 50 years old; 46% identify as Democrats and 36% as Republicans). We also oversampled members of the YouGov Pulse panel, who voluntarily provide behavioral data on their online information consumption ($N = 946$) (see Section B in the Supplementary Materials for demographic details). Pulse panelists confidentially share visit-level data on domains and URLs of web activity, including estimated duration and time stamps, on registered desktop/laptop and mobile devices. Data from Pulse panelists in our sample comprise $N = 11,903,134$ observations collected from laptop and desktop computers via the Reality Mine app. Secure transactions and passwords are not collected or shared with researchers, and YouGov performs a scrub of personally identifying information before delivering the data. For thorough validation of Pulse data, see Guess et al. (2020a); Guess (2021).

The main outcome of interest is news consumed by our study participants from publishers of low-quality news sources. Using Pulse data we measure this phenomenon with five different strategies for each respondent in three separate periods of the study: the period before they took the Wave 1 survey; the period between their completion of the Wave 1 survey and June 30;[20] and the two-week period after June 30, 2020. On July 1, NewsGuard transitioned from offering its extension for free to charging $2.99 a month. Given this switch, we assume that the vast majority of respondents were no longer using the NewsGuard extension after June 30. During these three periods, we calculate the average reliability score for websites visited, proportion of unreliable news sites visited, count of unreliable news sites visited, proportion of reliable news sites visited, and count of reliable news sites visited. The first measure, average reliability score, is derived by calculating the average NewsGuard reliability rating (0 to 100) of all news domains visited by the participant in that period (weighted by the number of page URLs visited in each domain). For the other measures, we labeled all news domains visited by that respondent in each period as unreliable (the domain has a reliability score from NewsGuard of below 60) or reliable (the domain has a reliability score from NewsGuard of 60 or above) and calculated the proportion and count of unreliable and reliable online news domains over the three periods of interest.

We are also interested in the effect of this intervention on the dependent variables specified in our other hypotheses and research questions, including the perceived accuracy of true and false news stories, trust in

[20]The NewsGuard extension's free capabilities terminated on June 30, 2020, coinciding with the end of the treatment period.

media, and other possible downstream effects. We are also interested if effects on these variables are higher within certain subgroups, such as those those who use social media more or that have lower levels of digital literacy. Details on all of these variables are available in Section A of the Supplementary Materials.

Treatment and Compliance

At the beginning of the Wave 1 survey, respondents were asked if they would be willing to install an extension to their web browser, which was intended to minimize differences between the treatment and control group and in compliance. We then randomly assigned respondents in Wave 1 to be encouraged to install the NewsGuard web extension. We do not find that those in the treatment and control groups were statistically different across income, race, partisanship, education, and gender (sample demographic details are presented in Section B of the Supplementary Materials). Those in the treatment group were slightly younger (by 2 years) and had slightly higher levels of digital literacy than the control group, but the magnitudes of these differences are small. We also found recontact rates in the treatment group and the control group to be similar (14.1% in the control group compared to 13.2% in the treatment group).[21] Based on Wave 2 survey data, 94% of participants in the treatment group who installed NewsGuard felt neutral or positive toward the extension, and 41% liked the extension "a little" or "a lot."

We define "compliance" as successfully installing and activating the NewsGuard extension (as a result of the encouragement), which we validate via a script linked at the beginning of the Wave 1 survey and during the last week of the treatment period.[22] This gives us two separate compliance measures that we can use for over 92% of our respondents.[23] Of those from whom we have data from, 95% passed the first compliance check and 80% passed both the first and second compliance check. Notably, we find little difference in the characteristics of respondents who would successfully take the treatment if encouraged ("compliers")[24] and those who would not take the treatment if encouraged ("never-takers"). We find no statistically significant evidence that respondents who comply are very much different than those who did not comply along the dimensions of age, partisanship, gender, or race. Compliers are more likely to hold

[21] We do not find significant differences in demographic characteristics between those in the control and treatment groups who did not complete Wave 2. We also do not find that treatment compliance rates differ between those who completed Wave 2 and those who did not (53% of respondents in the treatment group who attrited installed the NewsGuard extension, while 57% of respondents in the treatment group who did not attrite installed the extension). Details are presented in Section B of the Supplementary Materials.

[22] Respondents click the verification link in both compliance checks and they are redirected to a separate page in which we can verify whether the NewsGuard extension has been installed and is active. We record their unique ID and the result of the compliance check, so we can match it to their survey responses. We did not ask early respondents of our Wave 2 survey to complete the compliance check during the survey. Rather, we waited until the last week of the treatment period and sent them an e-mail asking them to click on the verification link. Respondents who filled out the Wave 2 survey in the last week of the treatment period were asked to click on the verification link at the end of their Wave 2 survey.

[23] This compliance check failed about 4% of the time due to random browser or survey issues that do not appear biased in any identifiable direction. Given this, we only collected first and second compliance check data for 92% of our respondents.

[24] In this analysis we define compliance using our second, stronger measure that uses the compliance checks during Wave 1 and during the last week of the treatment period.

a postsecondary degree, report a higher income level, and score higher on our digital literacy scale than never-takers, but the magnitudes of these differences are small.[25] Among the respondents for whom we have web data, compliers score higher on our digital literacy scale and consumed a higher proportion of unreliable news domains than never-takers in the pre-treatment period, but again, the magnitudes of these differences are small.

Analysis

Our pre-registered primary analyses are an Intent-to-Treat (ITT) model and two Complier Average Causal Effect (CACE) models using two different compliance measures, one that measures compliance solely at the beginning of the treatment period and another which measures compliance both at the beginning and end of the treatment period. We report both unadjusted (differences in means) and covariate-adjusted estimates of treatment effects for each dependent variable of interest.[26] For covariate-adjusted models, we selected covariates for inclusion using lasso regressions run separately for each dependent variable. The list of pre-treatment variables for possible inclusion as covariates can be found in Section A of the Supplementary Materials.

One concern with using an ITT model is that it will understate the true effect of an intervention when some respondents do not comply with the encouragement, but we report relatively high levels of compliance.[27] Given high levels of compliance we report the covariate-adjusted Intent-To-Treat effect in the main text of the paper and the Complier Average Causal Effect (CACE) using both measures of compliance in the Supplementary Materials.

References

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2).

Altay, S., Hacquin, A.-S., and Mercier, H. (2019). Why do so few people share fake news? it hurts their reputation. *new media & society*.

Anspach, N. M., Jennings, J. T., and Arceneaux, K. (2019). A little bit of knowledge: Facebook's news feed and self-perceptions of knowledge. *Research & Politics*, 6(1).

[25] Details comparing these two groups can be found in Section B of the Supplementary Materials.
[26] We use robust standard errors (HC2) in all analyses and report p-values from two-tailed t-tests.
[27] As discussed in the previous section, 95% of respondents passed the first compliance check and 80% passed both the first and second compliance check.

Bakshy, E., Messing, S., and Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Brennen, J. S., Simon, F., Howard, P. N., and Nielsen, R. K. (2020). Types, sources, and claims of covid-19 misinformation.

Chaiken, S. (1987). The heuristic model of persuasion. In *Social influence: the ontario symposium*, volume 5, pages 3–39.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., et al. (2019). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, pages 1–23.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Coppock, A., Ekins, E., Kirby, D., et al. (2018). The long-lasting effects of newspaper op-eds on public opinion. *Quarterly Journal of Political Science*, 13(1):59–87.

Eady, G., Nagler, J., Guess, A., Zilinsky, J., and Tucker, J. A. (2019). How many people live in political bubbles on social media? evidence from linked survey and twitter data. *Sage Open*, 9(1).

Ecker, U. K., Lewandowsky, S., and Tang, D. T. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition*, 38(8):1087–1100.

Flanagin, A. J. and Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, 77(3):515–540.

Fletcher, R., Cornia, A., and Nielsen, R. K. (2020). How polarized are online and offline news audiences? a comparative analysis of twelve countries. *The International Journal of Press/Politics*, 25(2):169–195.

Flynn, D., Nyhan, B., and Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38:127–150.

Gentzkow, M. and Shapiro, J. M. (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839.

Gerber, A. S., Gimpel, J. G., Green, D. P., and Shaw, D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? results from a randomized field experiment. *American Political Science Review*, pages 135–150.

Gigerenzer, G. and Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Gigerenzer, G. E., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior.* Oxford University Press.

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

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

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

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science*.

Guess, A. M., Barberá, P., Munzert, S., and Yang, J. (2021). The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14).

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

Hartman, E. and Hidalgo, F. D. (2018). An equivalence approach to balance and placebo tests. *American Journal of Political Science*, 62(4):1000–1013.

Hovland, C. I., Janis, I. L., and Kelley, H. H. (1953). Communication and persuasion.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Iyengar, S. and Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of communication*, 59(1):19–39.

Jurkowitz, M., Mitchell, A., Shearer, E., and Walker, M. (2020). U.s. media polarization and the 2020 election: A nation divided. Pew Research Center, January 24, 2020. Downloaded June 9, 2020 from `https://www.journalism.org/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/`.

Kim, A. and Dennis, A. R. (2019). Says who? the effects of presentation format and source rating on fake news in social media. *Mis Quarterly*, 43(3).

Kim, A., Moravec, P. L., and Dennis, A. R. (2019). Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems*, 36(3):931–968.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social psychological and personality science*, 8(4):355–362.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Lorenz-Spreen, P., Lewandowsky, S., Sunstein, C. R., and Hertwig, R. (2020). How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature Human Behaviour*, 0:1–8.

Lupia, A. (1994). Shortcuts versus encyclopedias: Information and voting behavior in california insurance reform elections. *American Political Science Review*, pages 63–76.

Messing, S. and Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research*, 41(8):1042–1063.

Munzert, S., Barberá, P., Guess, A., and Yang, J. (2021a). Do online voter guides empower citizens? evidence from a field experiment with digital trace data. *Public Opinion Quarterly*.

Munzert, S., Selb, P., Gohdes, A., Stoetzer, L. F., and Lowe, W. (2021b). Tracking and promoting the usage of a covid-19 contact tracing app. *Nature Human Behaviour*, pages 1–9.

Nyhan, B., Reifler, J., and Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical care*, pages 127–132.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., and Petersen, M. B. (Forthcoming). Partisan polarization is the primary psychological motivation behind 'fake news' sharing on twitter. *American Political Science Review*.

Paskhalis, T., Tucker, J., Bonneau, R., and Nagler, J. (2021). Path to fake: Who goes and how did they get there? *Unpublished Manuscript*.

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

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., and Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*.

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

Pennycook, G. and Rand, D. G. (2021). Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *PsyArXiv Preprints*.

Rainey, C. (2014). Arguing for a negligible effect. *American Journal of Political Science*, 58(4):1083–1091.

Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.