

# The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media

Tom Dobber, Sanne Kruike-meier, Fabio Votta, Natali Helberger, and Ellen P. Goodman

## ABSTRACT

The use of warning labels on political advertisements is one way to help citizens better evaluate the source and veracity of messaging, and combat the harms of misinformation on social media. Reliance on labeling is part of a larger policy push for greater transparency on social media platforms with respect to the source and quality of information. In this study, we test the effectiveness of “traffic light” labels (red, orange, green) as indicia of the veracity of political advertisements on YouTube. In an online experiment ( $N=1,054$ ), we test seven variations of TL-veracity labels and find that red and orange traffic light labels placed concurrently with the start of a political advertisement significantly affect credibility perceptions. Taken together, the findings suggest that direct-to-consumer labels can be effective inputs to credibility perceptions, but their effectiveness depends on timing and position.

## KEYWORDS

misinformation; disclosure; countermeasure; traffic lights; veracity labels; credibility perceptions


## How do traffic light labels help citizens better evaluate source and message veracity?

Mis- and disinformation in political advertising on social media produces harm, including misinforming citizens (Bakir & McStay, 2018; Tucker et al., 2018) and delegitimizing political institutions (Bennett & Livingston, 2018; Rid, 2020). This study uses misinformation as an umbrella term that refers to any information that is false (see Ecker et al., 2022). Especially in a polarized political environment, motivated reasoning may make people especially susceptible to false narratives in order to avoid contradiction (Flynn, Nyhan, & Reifler, 2017). Alternatively, research has suggested that a lack of reasoning makes people susceptible to false claims (Pennycook & Rand, 2018). Debiasing measures including fact-checks, source, and veracity labels are all ways to help citizens defend against misinformation (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012).

Measures to counter misinformation have shown promise in increasing citizens’ understanding of the information they are exposed to (e.g., Lewandowsky & van der Linden, 2021; Roozenbeek, van der Linden, & Nygren, 2020). Veracity labels are one such anti-measures. Veracity labels are a type of content label

that identifies the underlying claim as false or disputed, or may even directly correct the inaccuracies (Papakyriakopoulos & Goodman, 2022). Often based on independent fact-checking, veracity labels can help to “empower” end users, by providing citizens with a tool to evaluate the truthfulness of information (European Commission, 2020). Emerging regulatory strategies to combat misinformation call for more labeling. The draft EU regulation on transparency of political advertising, for example, explicitly requires the use of labeling techniques to comply with the Regulation’s transparency requirements (European Parliament, 2021). The European Code of Practice on Disinformation calls on signatories to embrace transparency measures that “reflect the importance of facilitating the assessment of content through indicators of the trustworthiness of content sources, media ownership and verified identity” (European Commission, 2018, p. 8). And the European Regulators Group for Audiovisual Media Services stated that labeling approaches are preferable to the removal of false content as a matter of fundamental rights, but also because labeling is more effective in raising awareness around the problem of misinformation (ERGA, 2021). Similarly, in the US, regulatory proposals to address online misinformation endorse

**CONTACT** Tom Dobber  [t.dobber@uva.nl](mailto:t.dobber@uva.nl)  Amsterdam School of Communication Research, University of Amsterdam, Nieuwe Achtergracht 166, Amsterdam 1018 WV, The Netherlands

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19331681.2023.2224316>.

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

labeling techniques (e.g., Consistent Labeling for Political Ads Act, 2021; Honest Ads Act, 2017).

Notwithstanding regulators' fondness for social media content labels, the emerging schemes focus very little on label positioning. Label effectiveness is contingent on design and positioning (see also Binford, Wojdyski, Lee, Sun, & Briscoe, 2021; van Drunen et al., 2022). First, some label designs make the messages more visible to people (Boerman & Kruikemeier, 2016; Kaiser et al., 2021; Wojdyski & Evans, 2016). Second, the messaging language can make the labels more understandable (Wojdyski & Evans, 2016). Third, proper positioning of the label increases the likelihood that people will pay attention (Wojdyski & Evans, 2016). Fourth, the timing of the label's appearance (before, during, or after a video advertisement) also affects the degree to which people will pay attention (van Reijmersdal et al., 2020). Despite the evidence that design and positioning affect label efficacy, there are currently no uniform evidence-based guidelines for label design and positioning. Social media platforms design labels that do not necessarily perform as intended (Binford, Wojdyski, Lee, Sun, & Briscoe, 2021) and are not harmonized. This lack of uniformity itself may be a source of user confusion, even apart from design deficiencies (ERGA, 2021).

This study sets out to test the efficacy of a particular kind of label – veracity labels that signal the truthfulness of a claim based on independent fact-checks. Studies show that people often fail to see or fail to register sponsorship disclosures (Amazeen & Wojdyski, 2019; Hoofnagle & Meleshinsky, 2015; Wojdyski & Evans, 2016). A label is more effective when users notice the label and it affects their assessment of the claim's truth. The study aims to learn from the nutrition literature, by testing veracity labels based on the colored traffic light labels used on food packages (Liu, Wisdom, Roberto, Liu, & Ubel, 2014). These traffic light labels have shown promise in encouraging healthier food choices (Thorndike, Riis, Sonnenberg, & Levy, 2014) and are thought to be intuitive and recognizable for citizens (Liu, Wisdom, Roberto, Liu, & Ubel, 2014). However, as of now, it is unclear to what extent user-facing traffic light veracity labels could be an effective means of helping citizens to better evaluate the

veracity of political information. By deploying an experimental design ( $N=1,054$ ), this study poses the following key question: To what extent are traffic light veracity labels effective in helping citizens evaluate the credibility of information in a political advertisement?

### ***Misinformation responses, different approaches***

There are two main de-biasing approaches to strengthen defenses to misinformation: pre-exposure interventions, such as inoculation (e.g., Roozenbeek, van der Linden, & Nygren, 2020) and nudges (e.g., Pennycook, Bear, Collins, & Rand, 2020), and post-exposure interventions, such as fact-checks. We take these in reverse order.

### ***Post exposure interventions: effects and limitations***

A meta-analysis of the evidence on the effectiveness of fact-checking in correcting political misinformation supports the supposition that fact-checking significantly influences beliefs (Walter, Cohen, Holbert, & Morag, 2020). Swire, Ecker, and Lewandowsky (2017) and Swire-Thompson, Ecker, Lewandowsky, and Berinsky (2020) find that corrections of false statements lead to a decreased belief in those corrected statements, and a higher belief in confirmed statements, which is in line with Hameleers and Van Der Meer (2020), and Nyhan, Porter, Reifler, and Wood (2020). There is some evidence that post-exposure debunking is more effective when combined with pre-exposure interventions. Hameleers (2020) experimentally tested the effectiveness of a combined media literacy intervention before exposure and a fact-check after exposure. Compared to the pre-exposure intervention only, and the post-exposure intervention only, the combined intervention was more effective in reducing misperceptions and perceived accuracy of misinformation. However, the efficacy of this integrative approach remains unclear. The findings of Hameleers (2020) are in line with Clayton et al. (2020), but Vraga, Bode, and Tully (2020) do not find evidence that a combined intervention is effective.

Post-exposure misinformation interventions are not always intuitive or easy to grasp. For instance,

fact-checks are often textual rebuttals of key claims made in news articles (e.g., Hameleers & van der Meer, 2020). Understanding these corrections requires cognitive effort and motivation. At times, fact checks can be so complex that even the more able reader might misinterpret their meaning (Nieminen & Sankari, 2021). In addition, post-exposure misinformation interventions may not reach all people that have been exposed to misinformation (Hameleers, 2020), and it remains unclear to what extent people's attitudes can be fully "corrected" with a post-exposure intervention. Indeed, Thorson (2016), as well as Walter and Tukachinsky (2020), find evidence that discredited information continues to influence people, *even* when the correction occurs directly after exposure.

### ***Pre-exposure: inoculation and persuasion knowledge***

Inoculation (McGuire, 1961) is an attempt to make people more resistant to persuasion ex ante (Roozenbeek, van der Linden, & Nygren, 2020). Inoculation occurs before the encounter with false information and consists of two elements. First, people are *forewarned* that they might be getting misinformed. Second, people get exposed to *non-harmful examples* of how they might be misinformed in the future (Lewandowsky & van der Linden, 2021). Thus, based on inoculation theory, it can be argued that a forewarning (such as a label prior to the subject content) triggers a "threat" that one might be vulnerable to a coming epistemic attack in the form of misinformation (Amazeen & Bucy, 2019). In other words, the label makes people more likely to be on guard for and, thus, recognize misinformation. Based on the persuasion knowledge model (Friestad & Wright, 1994), it can be argued that once misinformation is recognized, people will activate coping mechanisms that influence their attitudes, behavior, and knowledge. For instance, people might resist the persuasion attempt (Zuwerink Jacks & Cameron, 2003). Recent experimental work shows that inoculation can reduce susceptibility to misinformation (Roozenbeek, van der Linden, & Nygren, 2020), with effects stretching to two or three months after inoculation

(Maertens, Roozenbeek, Basol, & van der Linden, 2020). However, Zerback, Töpfl, and Knöpfle (2020) found limited and short-lived effects when they applied inoculation strategies to certain forms of online disinformation campaigns.

Warning labels are a form of pre-exposure intervention when they convey information about content *before* the content is consumed. In this study, we test a warning label for political communication that is often used to provide pre-consumption information about food: a traffic light labeling system (TLS). The health behavior literature suggests that traffic labels are more intuitive and easy to process than textual labels (Hawley et al., 2013). Our hypothesis is that graphic labels can improve upon textual labels with respect to user perception and cognitive processing of political advertising labels. In other words, we argue that especially traffic light labels are effective, because they are recognizable and their meaning is easily understood (Liu, Wisdom, Roberto, Liu, & Ubel, 2014).

The traditional, text-rich, way of presenting food quality information did not generally lead to healthier food choices (Scrinis & Parker, 2016). This is likely because consumers lack understanding or motivation to process such information (Liu, Wisdom, Roberto, Liu, & Ubel, 2014). To solve that problem, the UK Food Standards Agency (FSA) introduced traffic lights to label food. It is thought that traffic light graphics are especially intuitive because of the association people have with stop (red), caution (orange), and go (green; Liu, Wisdom, Roberto, Liu, & Ubel, 2014). The health behavior literature indicates that traffic light labels can be effective in increasing the awareness of food healthiness (Freire, Waters, Rivas-Mariño, Nguyen, & Rivas, 2017), and influencing behavior (Thorndike, Riis, Sonnenberg, & Levy, 2014). Exposure to TL-veracity labels in a political advertisement context is similarly expected to increase awareness about false information in the ad (see Freire, Waters, Rivas-Mariño, Nguyen, & Rivas, 2017). This is needed, because people in general do not consider accuracy when encountering information (Pennycook et al., 2021). Being made aware of the presence of false information, in turn, is likely to influence people's credibility perceptions (similar to Pennycook et al., 2021).

Where TL-veracity labels excel in ease of understanding, it has the drawback of conveying minimal amounts of information, making such labels useful only for signaling very simple messages. In the misinformation space, this means that TL-veracity labels are only appropriate when the truth or falsity of the underlying communication is relatively uncontroversial.

### **Message accuracy perceptions**

Pro-active, and substantially low-effort *nudges* have recently been shown effective in combatting misinformation. Pennycook et al (2021, 2020) found that a relatively simple nudge to help people consider the accuracy of news stories reduced the sharing (intentions) with respect to low-quality news stories on social media. A replication of Pennycook, Bear, Collins, and Rand (2020), by Roozenbeek, Freeman, and van der Linden (2021) also found a similar effect, albeit a much smaller one that wears off quickly (after about seven headlines).

Traffic light labels are also pro-active, low-effort nudges. In comparison with the *accuracy nudges* used by Pennycook et al (2021, 2020). TL-veracity labels are more paternalistic and more intrusive because TL-veracity labels are related to specific pieces of information rather than the higher-level concept of accuracy of information in general. However, like Pennycook et al (2021, 2020), as well as the TL food literature (Freire, Waters, Rivas-Mariño, Nguyen, & Rivas, 2017), this study expects that the TL-veracity label interventions can influence the perceptions of the accuracy of the message.

### **Source credibility**

In addition to the accuracy perceptions of the message, it is important to consider the credibility perceptions of the misinformation *source*. Credible sources spread misinformation more effectively than non-credible sources (Pluviano, Della Sala, & Watt, 2020; Swire, Ecker, & Lewandowsky, 2017). Moreover, corrections are less effective if the misinformation is perceived to have originated from a credible source (Swire, Ecker, & Lewandowsky, 2017). Aslett, Guess, Bonneau, Nagler, and Tucker (2022) found that

prolonged exposure to color-coded source credibility cues did not decrease consumption of news from unreliable sources, nor did it increase trust in mainstream media or reliable news sources. Aslett, Guess, Bonneau, Nagler, and Tucker (2022), findings suggest source-credibility interventions increased the news diet quality “among the heaviest consumers of misinformation” (p. 7). However, effects might be underestimated due to the unobtrusive design of the intervention and because participants were only modestly exposed to low-credibility sources.

This study argues that lowering or increasing source credibility perceptions should be a direct goal of misinformation interventions in the political advertising context. Elite actors play an important role in the spread of misinformation, precisely because they are often perceived as credible sources. Van Duyn and Collier (2019) found that elite discourse about “fake news” has the potential to decrease people’s trust in the media and even hamper people’s ability to identify real news. Taken together, a TLS intervention casting doubt on the veracity of the information might directly decrease source credibility. The TLS labels should help the user to attribute credibility to the source of the video ad, before (or concurrent with) exposure. This leads to the following hypotheses.

**H1a:** Exposure to a red TL-veracity label decreases perceived source credibility and message credibility compared to no exposure to a label.

**H1b:** Exposure to an orange TL-veracity label decreases perceived source credibility and message credibility compared to no exposure to a label.

### **Green light veracity label**

The above section assumes that the user sees a red or orange TL-veracity label. However, a TLS intervention also suggests a *green light*, which means that the veracity of the information is not in question. Displaying a green light creates a different dynamic because, other than the red and orange light, the green light is not meant to decrease accuracy perceptions, source credibility, or engagement intentions. Indeed, the green light serves as

a “stamp of approval,” signaling to citizens in an intuitive way that they can trust the information. Pennycook, Bear, Collins, and Rand (2020) find that people were more likely to consider sharing headlines when those headlines were tagged with large notable “TRUE” labels. Based on that, this study expects the following.

**H1c:** Exposure to a green TL-veracity label increases perceived source credibility and message credibility, compared to no exposure to a label.

### **Labeling unchecked information**

The implied truth effect means that people who have seen misinformation warnings perceive messages without such warnings as more accurate (Pennycook, Bear, Collins, & Rand, 2020). When applying the implied truth effect to TL-veracity labels, it can be expected that the absence of a label could be taken as a cue that the information must be true. To counter the implied truth effect, unchecked content could be labeled with a “veracity-not-yet-established” message so people know that the absence of a fact check label does not imply truth. However, such a label could make people more skeptical of all information, regardless of its veracity. This leads to the following hypothesis.

**H1d:** Exposure to a “veracity-not-yet-established”-label decreases perceived source credibility and message credibility compared to no exposure to a label, but to a lower extent than a red or orange TL-veracity label.

### **Timing of the label**

Research focusing on sponsorship disclosures of social influencers, it has been found repeatedly that disclosures shown before the start of social influencers’ videos are more effective than disclosures shown concurrent with the start of the video in increasing visual attention (van Reijmersdal et al., 2020) and cognitive advertising literacy (De

Pauw, Hudders, & Cauberghe, 2018). It is thought that showing a disclosure before the start of the video enables viewers to better process the information conveyed by the label (van Reijmersdal et al., 2020). However, these studies focused on children. Research on adults found differential effects of timing. Campbell, Mohr, and Verlegh (2013) found that prior disclosures affected only recall, while corrections *afterward* affected both recall and attitudes. On the other hand, Boerman, van Reijmersdal, and Neijens (2014) found that sponsorship disclosures shown prior to or concurrent with the sponsored content increased sponsorship recognition, while disclosures shown afterward were not effective. Literature on misinformation corrections is not unequivocal in terms of timing either. Some studies suggest debunking is more effective than prebunking or labeling (e.g., Brashier, Pennycook, Berinsky, & Rand, 2021), other suggest that prebunking is more effective than debunking (e.g., Jolley & Douglas, 2017). However, the misinformation-correction literature often focuses on news headlines or articles rather than on political advertisements. Based on the contradictory nature of the literature, we formulate the following research question.

**RQ1:** To what extent are there differences in effectiveness between TL-veracity labels shown before or concurrent with the start of the video advertisement?

### **False information awareness**

People often fail to consider information quality when sharing information (Mosleh, Pennycook, Arechar, & Rand, 2021). Pennycook et al. (2021) found that making people aware of the concept of accuracy before sharing information led people to spread less misinformation. Like the field experiment in Pennycook et al. (2021), this study exposes people to an information quality nudge. Different from Pennycook et al. (2021), this study uses TLS labels in a YouTube political ad environment rather than direct messages in a Twitter news headlines environment. However, the mechanism should be similar: TLS labels, and to a lesser degree the “veracity not-yet-established

label” should nudge people into becoming aware of information quality. However, many studies show that labels, or other types of warnings, are often unnoticed (Boerman & Kruikemeier, 2016) and people might not understand their meaning (i.e., externalize threat). Especially in an online environment, people might become blind to labels and consequently do not see, understand, or process them (Benway, 1998; Burke, Hornof, Nilsen, & Gorman, 2005). As we mentioned in the previous section, we expect that people will only activate persuasion knowledge or resist a message when they understand that the label implies forthcoming misinformation. In other words, an important condition under which labels affect the credibility of the source and message is the extent to which the message makes people aware of false information. Thus, when people are aware of being exposed to false information, this will affect source credibility and message credibility in the following way.

**H2a:** Exposure to a red TL-veracity label leads to increased awareness of false information compared to no label, and this, in turn, negatively affects source credibility and message credibility.

**H2b:** Exposure to an orange TL-veracity label leads to increased awareness of false information compared to no label, and this, in turn, negatively affects source credibility and message credibility.

**H2c:** Exposure to a green TL-veracity label leads to decreased awareness of false information compared to no label, and this, in turn, positively affects source credibility and message credibility.

**H2d:** Exposure to a veracity not-yet-established label leads to increased awareness of false information compared to no label, and this, in turn, negatively affects source credibility and message credibility but to a lower extent than a red or orange TL-veracity label.

## Method

Participants were recruited by Kantar Lightspeed, a Dutch company that specialized in recruitment

for academic purposes, and were paid a small amount for their participation. Data collection took place in October 2019. The sample consisted of 1,550 participants. We removed speedsters and participants who did not view the stimulus in its entirety ( $N = 454$ ). We also report a small number of missing values for our control variables, age, gender, and political interest ( $N = 42$ ). This left us with a total of 1,054 participants. The mean age in the sample was 44.12 ( $SD = 13.45$ ), and 53.51% were female. The mean political interest in the sample was 4.48 ( $SD = 1.48$ ), on a scale from 1 to 7 (7=higher level of political interest). The typical participant finished the online experiment within 5 to 10 minutes. All participants were debriefed.

## Design

This experiment consisted of seven treatment conditions and one control condition. Participants were randomly placed into one of those conditions.

## Independent variables

Participants saw the same 46 second video advertisement calling for an increase in the minimum wage. In Dutch politics, this is an uncontroversial and broadly supported issue (I&O Research, 2020). This political issue ad was sponsored by Dutch union FNV, but references to FNV were removed from the video. People in the control condition saw only this video. But the participants in the experimental conditions were exposed to one of seven variations of traffic light interventions (see Appendix A). First, there was a red traffic light, accompanied by a text in Dutch stating “this message contains false information.” Second, there was an orange traffic light, accompanied by the text “this message contains partly false information.” Third, there was a green traffic light, with the text “this message contains no false information.” These three stimuli were shown for 10 seconds *before* the video advertisement started. Fourth, there was a red traffic light, accompanied by the text “this message contains false information” shown in the top right corner, concurrent with the start of the video and also shown for 10 seconds. Five, there also was a similar orange traffic light variant. Six, there was also a green variant. Seven, there was

a disclaimer text (white letters against a black background) stating that in Dutch that “the veracity of the information in this video has not yet been checked.” This text was shown *before* the video.

*False information awareness* is the mediator and was measured on a 7-point scale. The variable consisted of the following two items: *the video contained false information*, and *the video is fake* ( $r = .70$ ;  $M = 3.73$ ;  $SD = 1.62$ ).

*Source credibility* was measured with the following 11 items, all on 7-point scales (7 is higher credibility): *Based on this video, I think the source of this video is . . . trustworthy, experienced, honest, knowledgeable, authentic, sincere, competent, has expertise on the matter, capable, credible, appealing*. Together, these items formed a scale ( $M = 4.90$ ;  $SD = 1.30$ ), with an Eigenvalue of 8.35. The Cronbach’s alpha was 0.97.

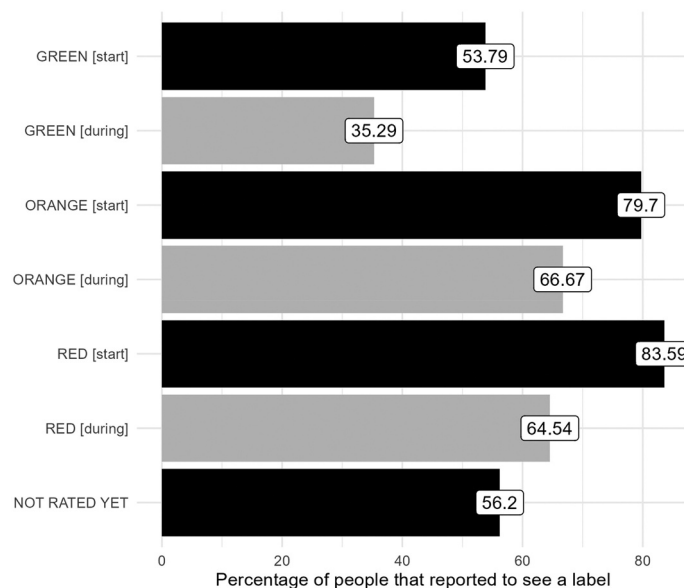
*Message credibility* was measured with the following eight items, all on 7 point scales (7 is higher credibility): *I can trust that this video tells the truth; I think the goal of this video is to inform the Dutch citizens; I find the video informative; I find the video sincere; The video is a reliable source of information; The video reflects the truth well; I feel like I have been informed correctly by the video; In general, the video creates a reliable picture of the FNV*. Together, these items formed a scale ( $M = 4.38$ ;

$SD = 1.52$ ), with an Eigenvalue of 6.34. The Cronbach’s alpha was 0.97.

*Political interest* was measured using the following three items, all on 7-point scale (7 is higher interest). *How much interest do you have in political issues in general; in local politics (for example the politics in your residential municipality); in national politics?* Together, these items formed a scale ( $M = 4.49$ ;  $SD = 1.50$ ; Eigenvalue = 2.16; Cronbach’s alpha = .90).

## Checks

After measuring the relevant variables, we asked participants whether they had seen a warning with information about false information in the video. Participants could answer yes, or no. [Figure 1](#) shows per condition how many participants noticed the stimulus. The participants who falsely remembered not seeing the stimulus are included in our analyses because not remembering correctly does not automatically mean not seeing the stimulus. However, in [Appendix C](#), we report the analyses on the smaller group of participants who correctly remembered seeing the stimulus ( $N = 710$ ). In the next section, at the end of each hypothesis test, we will indicate whether this finding holds



**Figure 1.** Overview of participants who noticed the TL-label, per condition. Note that there are 24 participants in the control condition who reported to have seen a label, even though there was no label treatment. Excluding them from the control condition does not impact the findings of the analysis so we decided to keep them.

when we look only at the group of people who correctly remembered seeing the stimulus.

Ethical approval was granted by the ethics board of (name withheld for peer review), and was given the following approval number (withheld for peer review).

## Results

### Source credibility

We fit linear models (estimated using OLS) with the dependent variable source credibility and the various treatment conditions as independent variables (reference category is the control condition). Model 1 shows a regression with only control variables (age, gender, and political interest) and Model 2 adds the seven treatment conditions with the reference category being the control condition.

In the following, we will focus our interpretations on the more complete Model 2 ( $F(10, 1043) = 7.86$ ,  $p < .001$ ; adj.  $R^2 = .06$ ). Figure 2 plots the regression coefficients for Model 2 in blue (Table SB1 in Appendix B shows regression results for Models 1 and 2).

First, contrary to what was expected by H1a, showing a red light concurrent with the video does not significantly reduce source credibility compared to showing no label ( $b = -.25$ , 95% CI  $[-0.55, 0.05]$ ,  $t(1043) = -1.64$ ,  $p = .102$ ). The effect of showing an orange light concurrent with the video on source credibility is statistically significant and negative, as expected by H1b ( $b = -0.60$ , 95% CI  $[-0.90, -0.29]$ ,  $t(1043) = -3.84$ ,  $p < .001$ ). However, contrary to expectations, the red and orange light counterparts that were shown *before* the video started had no discernable effect on source credibility compared to the control

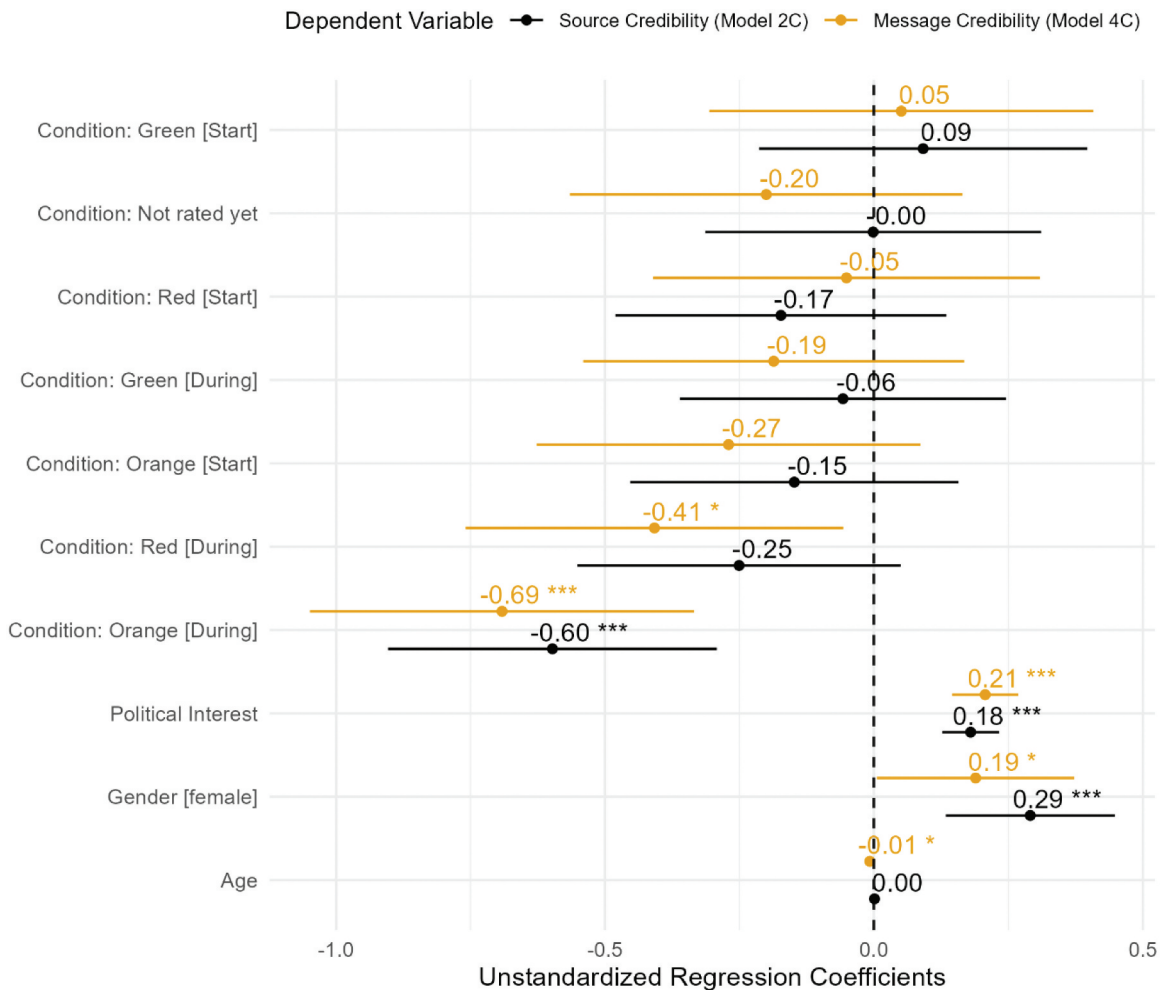


Figure 2. Regression coefficients from Model 2 and 4. \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$



condition (see Figure 2). This means that hypothesis 1b is partially supported. Only the orange-light label shown concurrent with the start of the video significantly reduces source credibility perceptions.

### **Message credibility**

We fit linear models (estimated using OLS) with the dependent variable message credibility and the various treatment conditions as independent variables (reference category is the control condition). Model 3 shows a regression with only control variables (age, gender, and political interest) and Model 4 adds the seven treatment conditions with the reference category being the control condition. In the following, we will focus our interpretations on the more complete Model 4 ( $F(10, 1043) = 7.62$ ,  $p < .001$ ; adj.  $R^2 = .06$ ). Figure 2 plots the regression coefficients for Model 4 in red (Table SB2 in Appendix B shows regression results for both Model 3 and 4).

First, we can observe that the results for message credibility are slightly different from source credibility: red or orange lights concurrent with the video were the only interventions to have any statistically discernible effects on message credibility. As expected by H1a, showing a red light concurrent with the video reduces message credibility by 0.41 scale points compared to no label ( $b = -.41$ , 95% CI  $[-.76, -.06]$ ,  $t(1043) = -2.28$ ,  $p = .02$ ). Similarly, the effect of showing an orange light concurrent with the video on source credibility is statistically significant and negative, as expected by H1b ( $b = -.69$ , 95% CI  $[-1.05, -.33]$ ,  $t(1043) = -3.80$ ,  $p < .001$ ). This means that H1a and H1b are partially supported by the data. Indeed, an orange-light label decreases source and message credibility perceptions, but only if the label is shown concurrent with the start of the video. The red TL-veracity label significantly decreases message credibility but not source credibility perceptions, provided that the label is shown concurrent with the video. Exposure to a green-light label or a veracity-not-yet-established label had no significant effect on source or message credibility perceptions. Hypotheses 1c and 1d are not supported. When running the same analysis for the ATT group (the smaller group of participants that correctly remembered seeing the stimulus;  $N = 710$ ), we see

a similar picture. Only the orange and red TL-veracity labels shown concurrent with the start of the video decrease source credibility and message credibility significantly (see Appendix C). This is slightly different from the larger group of participants, for which the red TL-veracity label shown concurrent with the start of the video had no significant effect on source credibility.

### **Timing of the label**

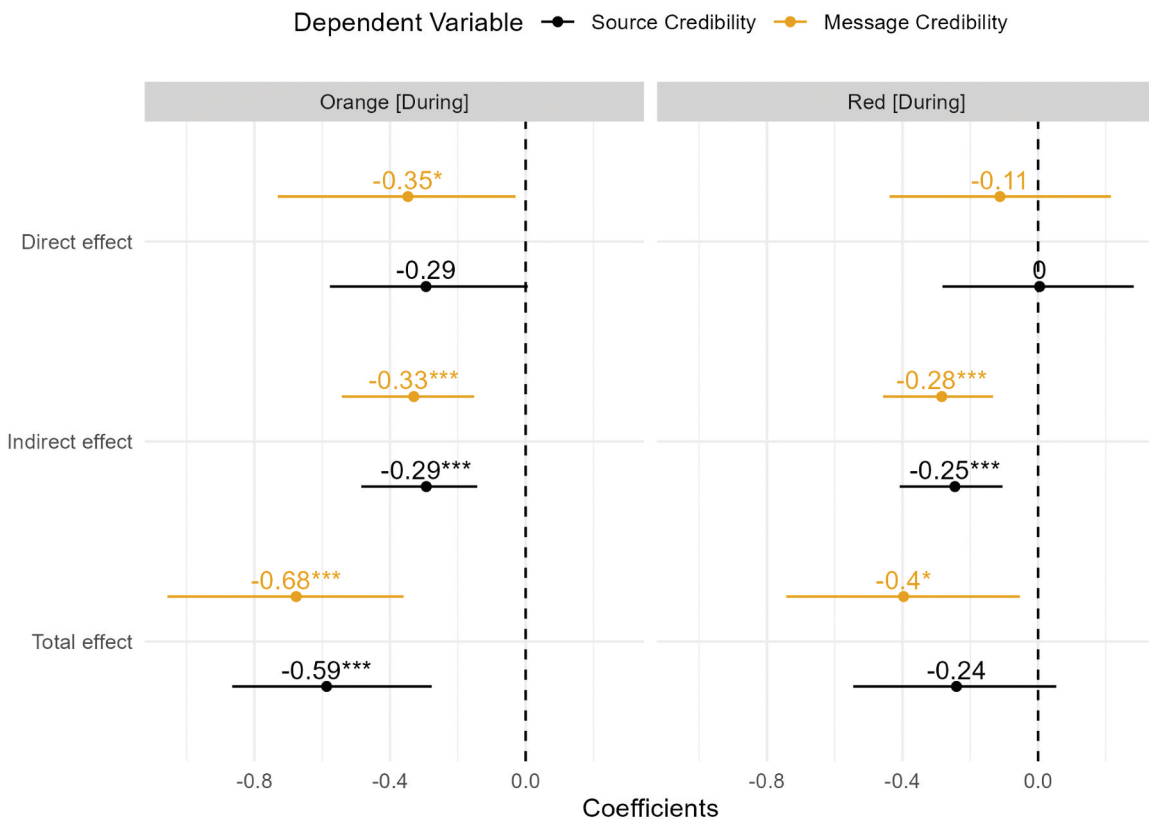
RQ1 asked to what extent there are differences in effectiveness between TL-veracity labels shown before or concurrent with the start of the video advertisement. Figure 2 shows that only red and orange traffic light labels shown concurrent with the start of the video are effective in decreasing credibility perceptions.

### **Mediators**

We only report mediator effects of false information awareness for the treatment conditions that had statistically significant direct effects: red and orange-light labels during the start of the video. Because we found no significant direct effects for the green TL-veracity label and the veracity-not-yet-established label, hypotheses 2c and 2d are not supported by the data. We do find that the effects on message credibility of the orange and red lights shown concurrent with the video were completely mediated by false information awareness. The effect on source credibility of the orange light shown concurrent with the video was partially mediated through false information awareness.

### **Mediated effect of orange-light label on source and message credibility**

First, we look at the dependent variable source credibility and the orange-light label shown concurrently with the video. The independent variable used here is the binary condition between “orange light during video” vs. “control condition.” The mediating variable is false information awareness, and we also use the control variables age, gender, and political interest. The results are shown in blue in the right panel of Figure 3.



**Figure 3.** Mediation effects for treatment conditions red during and orange during. Indirect effect is effect of exposure to TL-veracity label, mediated by false information awareness, on source credibility (blue) or message credibility (red). \*\*\*  $p < 0.001$ , \*\*  $p < 0.01$ , \*  $p < 0.05$

Total effect on source credibility is  $-.59$  and statistically significant (total effect =  $-.59$ , 95% CI  $[-.87, -.28]$ ,  $p < .001$ ). The direct effect of an orange light on source credibility removing false information awareness is not statistically significant (direct effect =  $-.29$ , 95% CI  $[-.58, .01]$ ,  $p > .05$ ). The indirect effect of orange light on source credibility through false information awareness however is statistically significant (indirect effect =  $-.29$ , 95% CI  $[-.48, -.014]$ ,  $p < .001$ ). These findings indicate that the effect on source credibility of the orange light shown concurrent with the start of the video is completely mediated through false information awareness and offer partial support for hypothesis H2b.

Next, we take a look at the mediated effect of the orange light on message credibility. The results are shown in red in the left panel of Figure 3. Total effect on message credibility is  $-.68$  and statistically significant (total effect =  $-.68$ , 95% CI  $[-1.06, -.36]$ ,  $p < .001$ ). The direct effect of an orange light on message credibility removing false information awareness is statistically significant (direct

effect =  $-.35$ , 95% CI  $[-.73, -.03]$ ,  $p < .05$ ). The indirect effect of an orange light on message credibility through false information awareness is statistically significant (indirect effect =  $-.33$ , 95% CI  $[-.54, -.15]$ ,  $p < .001$ ). In sum, these findings show that the effect on message credibility of the orange light shown concurrent with the start of the video is partially mediated through false information awareness, and offer partial support for hypothesis H2b.

### Mediated effect of red-light label on message credibility

Next, we turn to the red-light label shown concurrently with the video. We found a direct effect of the red-light label on message credibility but not source credibility. This, we focus only on the mediated effect of the red light on message credibility. The independent variable used here is the binary condition between “red light during video” vs. “control condition.” The mediating variable is

false information awareness, and we also use the control variables age, gender, and political interest. The results are shown in red in the right panel of [Figure 3](#). Total effect of red light on message credibility is  $-.40$  and statistically significant (total effect =  $-.40$ , 95% CI [ $-.74, -.05$ ],  $p < .05$ ). The direct effect of a red light on message credibility removing false information awareness is not statistically significant (direct effect =  $-.11$ , 95% CI [ $-.44, .21$ ],  $p > .05$ ). The indirect effect of a red light on message credibility through false information awareness is statistically significant (indirect effect =  $-.28$ , 95% CI [ $-.46, -.13$ ],  $p < .001$ ). These findings indicate that the effect on message credibility of the red light shown concurrent with the start of the video is completely mediated through false information awareness and offer partial support for hypothesis H2a.

## Discussion

In this study, we set out to learn from the health behavior literature and test the extent to which the use of traffic light interventions against false or partly false information could decrease credibility perceptions of misinformation. In line with the health behavior literature (Freire, Waters, Rivas-Mariño, Nguyen, & Rivas, 2017; Thorndike, Riis, Sonnenberg, & Levy, 2014), we found that there is some promise in traffic light warnings.

At the same time, and in line with the integrative misinformation correction literature (see Clayton et al. (2020); Hameleers (2020), Vraga, Bode, and Tully (2020), our findings are not unambiguous. This study finds that the red and orange lights shown concurrently with the start of the video are effective in lowering perceptions of credibility, but all the other interventions did not significantly affect citizens' perceptions. In line with Wojdyski and Evans (2016) and van Reijmersdal et al. (2020), Campbell, Mohr, and Verlegh (2013), Boerman, van Reijmersdal, and Neijens (2014), Brashier, Pennycook, Berinsky, and Rand (2021), Jolley and Douglas (2017) this suggests that positioning as well as timing matter greatly.

The literature on the timing of labels and disclosures is contradictory. Our finding that only TL-veracity labels shown *concurrent* with the start of the video are effective in lowering credibility

perceptions, while labels shown prior to the video are not, does not add more clarity to the disclosure literature. Indeed, the finding is directly at odds with van Reijmersdal et al. (2020), De Pauw, Hudders, and Cauberghe (2018) who found that prior disclosures were most effective. The finding also partly contradicts Boerman, van Reijmersdal, and Neijens (2014), who found that disclosures prior to as well as concurrent with the start of the advertisement were effective. The finding also contradicts Brashier, Pennycook, Berinsky, and Rand (2021), who found debunking more effective than prebunking and labeling, and contradicts Jolley and Douglas (2017), who found prebunking most effective. It is clear that there is no consensus yet on the ideal timing. Potentially this is due to variations within a population, or this interacts with the design of the intervention (e.g., the topic: minimum wage). It must be noted that van Reijmersdal et al. (2020) focused on commercial native advertising in social influencer videos. Potentially the inconspicuousness of native advertising (in comparison with regular advertising) is a factor that future research could explore.

The finding that only labels shown concurrent with the video are effective in lowering credibility perceptions might partly be explained by reactance or by banner blindness. Persuasion literature has found that forewarning might lead to reactance, and more elaborate forewarning especially does so (Richards, Banas, & Magid, 2017). [Figure 1](#) shows that the prior label is noticed more often than the label shown concurrently. People might have experienced the more prominent prior label as too directive, which then produced reactance (Brehm, 1966). Banner blindness might also play a role. Banner blindness occurs when people do not remember noticing, or even avoid noticing online (banner) advertisements (Hervet, Guérard, Tremblay, & Chtourou, 2011; Owens, Chaparro, & Palmer, 2011). Indeed, Burke, Hornof, Nilsen, and Gorman (2005) found that banner blindness leads people to not see, understand, process, or remember seeing banners. Users confronted with a TL-veracity label before the start of the video might mistake the label for an advertisement and thus do not spend cognitive resources to the label.

In a similar vein, there were relatively large groups of people that did not notice the TLS label. The green-light label scored rather low

(46% did not notice label shown at start; 65% did not notice label shown concurrently; see [Figure 1](#)). This is possibly a measurement error. The participants were asked whether they had seen a label about false information in the video. Those in the green condition saw a green traffic light with the text: “this advertisement does not contain false information.” While participants technically saw a label about false information, this way of asking is likely too confusing. It could also be that people tend not to invest intellectual resources in a label that signals that there is no problem. This would suggest that the use of green TLS labels as a way of encouraging the consumption/sharing of accurate information might be less effective than labels that seek to discourage the consumption/sharing of inaccurate information. This would be relevant to the Code of Practice on Disinformation (European Commission, 2018), which explicitly calls on signatories to invest in measures to ensure the findability of trustworthy content. If people do not see green-light labels, labeling might be a less effective tool.

TL-veracity labels can convey only minimal amounts of information, making such labels most useful for signaling very simple messages. In the misinformation space, this means that TL-veracity labels are most appropriate when the truth or falsity of the underlying communication is relatively uncontroversial. As misinformation per definition is spread unintentionally, TL-veracity labels are especially useful for misinformation. Disinformation, however is spread intentionally, often around polarizing issues (Bennett & Livingston, 2018). The falsity of disinformation is thus typically more controversial and perceptions of falsity of the underlying communication has been found to be solidified by people’s (political) identity (Ecker et al., 2022; Van Duyn & Collier (2019)) and confirmation bias (Hameleers and Van der Meer, 2022). Thus, TL-veracity labels are less likely to affect people’s accuracy estimations around disinformation, and are more suited for misinformation. It is important to note that this study strictly measures effects of the TL-veracity labels. However, while the stimuli warned about false information in the video, the videos did not actually contain false information. Moreover, the videos were about minimum wage, which is

a broadly supported issue in the Netherlands. As most disinformation is about divisive topics, future research should explore whether TL-veracity labels are equally effective when applied to divisive disinformation.

We found that specific types of TL-veracity labels were able to lower source and message credibility perceptions. Research shows that source and message credibility perceptions can mediate accuracy perceptions of misinformation (e.g., Traberg & van der Linden, 2022; Kim et al., 2020), as well as sharing intentions (Ali, Li, Zain-Ul-Abdin, & Zaffar, 2021). As such, decreasing source and message credibility perceptions is a crucial step in efforts to prevent misperceptions following misinformation. This current study shows that, at an earlier stage, source and message credibility perceptions are mediated by “false information awareness”. This insight furthers our understanding of the mechanism behind the prevention and correction of misinformation-induced misperceptions. Arguably, future interventions should especially be focused on lowering message credibility. While source credibility is also important, message credibility is a more practical target for veracity labels because: (1) the message itself contains the misinformation that needs to be warned against; and (2) people or organizations can spread misinformation without in the first instance being non-credible sources. For instance, the New York Times sometimes “gets it wrong”, but that doesn’t make the newspaper a non-credible source. Similarly, a political institution could make a mistake in its communication, but that should not discredit that political institution. It becomes more problematic when people or organizations repeatedly, perhaps willingly, spread false information. In these cases, the messages may well qualify as disinformation and that means that the success of warnings is more dependent on people’s political identity (Ecker et al., 2022; Van Duyn & Collier, 2019) and confirmation bias (Hameleers and Van der Meer, 2020), or that we are potentially facing Internet Research Agency actors who have many different disposable sources (Linville & Warren, 2020).

Overall, the traffic light veracity label shows some promise in decreasing credibility perceptions, but label positioning and timing matter. Current

regulatory proposals, such as the Political Advertising Regulation (European Parliament, 2021), do require the use of labeling techniques, but are unspecific about how these labels should look, or when these labels should be shown to users (see van Drunen et al., 2022). This study shows that if lawmakers want to implement labels as a tool to empower citizens, they should not leave the positioning and timing of such labels to the social platforms but rather base these choices on further empirical research.

Another direction of further research could be to examine whether it matters what text accompanies the label. Or whether it matters for people to know who fact-checked (e.g., an independent fact checker, or the platform). Finally, research could focus on which conditions would need to be fulfilled for an effective labeling approach (think of issues such as standardization, public communication, self-regulation versus coregulation or formal regulation, but also of the question whether labels work better for some people than for others).

### Disclosure statement

No potential conflict of interest was reported by the authors.

### Funding

The work was supported by the John S. and James L. Knight Foundation (GR: -2019-59641), the European Research Council (ERC: 949754), and NWO Grant MVI.19.019.

### Notes on contributors

*Dr. Tom Dobber* is assistant professor at the Amsterdam School of Communication Research (ASCoR), University of Amsterdam.

*Prof. dr. Sanne Kruikemeier* is full professor Strategic Communication at Wageningen University.

*Prof. dr. Natali Helberger* is University Professor Law & Digital Technology, with a special focus on AI, University of Amsterdam.

*Prof. dr. Ellen P. Goodman* is Professor of Law at Rutgers Law School and the co-director and co-founder of the Rutgers Institute for Information Policy & Law.

### References

- Ali, K., Li, C., Zain-Ul-Abdin, K., & Zaffar, M. A. (2021). Fake news on Facebook: Examining the impact of heuristic cues on perceived credibility and sharing intention. *Internet Research*, 32(1), 379–397. doi:10.1108/INTR-10-2019-0442
- Amazeen, M. A., & Bucy, E. P. (2019). Conferring resistance to digital disinformation: The inoculating influence of procedural news knowledge. *Journal of Broadcasting and Electronic Media*, 63(3), 415–432. doi:10.1080/08838151.2019.1653101
- Amazeen, M. A., & Wojdyski, B. W. (2019). Reducing native advertising deception: Revisiting the antecedents and consequences of persuasion knowledge in digital news contexts. *Mass Communication and Society*, 22(2), 222–247. doi:10.1080/15205436.2018.1530792
- Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*, 8(18), eabl3844. doi:10.1126/sciadv.abl3844
- Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, 6(2), 154–175. doi:10.1080/21670811.2017.1345645
- Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, 33(2), 122–139. doi:10.1177/0267323118760317
- Benway, J. P. (1998). Banner blindness: The irony of attention grabbing on the World wide web. *Proceedings of the Human Factors and Ergonomics Society*, 1, 463–467. doi:10.1177/154193129804200504
- Binford, M. T., Wojdyski, B. W., Lee, Y. I., Sun, S., & Briscoe, A. (2021). Invisible transparency: Visual attention to disclosures and source recognition in Facebook political advertising. *Journal of Information Technology & Politics*, 18(1), 70–83. doi:10.1080/19331681.2020.1805388
- Boerman, S. C., & Kruikemeier, S. (2016). Consumer responses to promoted tweets sent by brands and political parties. *Computers in Human Behavior*, 65, 285–294. doi:10.1016/j.chb.2016.08.033
- Boerman, S. C., van Reijmersdal, E. A., & Neijens, P. C. (2014). Effects of sponsorship disclosure timing on the processing of sponsored content: A study on the effectiveness of European disclosure regulations. *Psychology & Marketing*, 31(3), 214–2249. doi:10.1002/mar.20688
- Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences of the United States of America*, 118(5), 2–4. doi:10.1073/pnas.2020043118
- Brehm, J. W. (1966). *A theory of psychological reactance*. New York: Academic Press.

- Burke, M., Hornof, A., Nilsen, E., & Gorman, N. (2005). High-cost banner blindness: Ads increase perceived workload, hinder visual search, and are forgotten. *ACM Transactions on Computer-Human Interaction*, 12(4), 423–445. doi:10.1145/1121112.1121116
- Campbell, M. C., Mohr, G. S., & Verlegh, P. W. J. (2013). Can disclosures lead consumers to resist covert persuasion? The important roles of disclosure timing and type of response. *Journal of Consumer Psychology*, 23(4), 483–495. doi:10.1016/j.jcps.2012.10.012
- Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., & Nyhan, B. (2020). 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*, 42(4), 1073–1095. doi:10.1007/s11109-019-09533-0
- Consistent Labeling for Political Ads Act. (2021). *US congress*. <https://www.congress.gov/bill/117th-congress/house-bill/989>
- De Pauw, P., Hudders, L., & Cauberghe, V. (2018). Disclosing brand placement to young children. *International Journal of Advertising*, 37(4), 508–525. doi:10.1080/02650487.2017.1335040
- Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K. . . . Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1), 13–29. doi:10.1038/s44159-021-00006-y
- ERGA. (2021). *Strengthening factchecking across the European Union (ERGA report)*. Brussels: ERGA.
- European Commission. (2018). *Tackling online disinformation, a European approach*. Brussels: European Commission.
- European Commission. (2020). *Democracy action plan (DAP)*. Brussels: European Commission.
- European Parliament. (2021). *Proposal on the transparency and targeting of political advertising*. Brussels: European Parliament.
- Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, 38, 127–150. doi:10.1111/pops.12394
- Freire, W. B., Waters, W. F., Rivas-Mariño, G., Nguyen, T., & Rivas, P. (2017). A qualitative study of consumer perceptions and use of traffic light food labelling in Ecuador. *Public Health Nutrition*, 20(5), 805–813. doi:10.1017/S1368980016002457
- Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research*, 21(1), 1. doi:10.1086/209380
- Hameleers, M. (2020). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information Communication and Society*, 25(1), 110–126. doi:10.1080/1369118X.2020.1764603
- Hameleers, M., Powell, T. E., Van Der Meer, T. G. L. A., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, 37(2), 281–301. doi:10.1080/10584609.2019.1674979
- Hameleers, M., & van der Meer, T. G. L. A. (2020). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, 47(2), 227–250. doi:10.1177/0093650218819671
- Hawley, K. L., Roberto, C. A., Bragg, M. A., Liu, P. J., Schwartz, M. B., & Brownell, K. D. (2013). The science on front-of-package food labels. *Public Health Nutrition*, 16(3), 430–439. doi:10.1017/S1368980012000754
- Hervet, G., Guérard, K., Tremblay, S., & Chtourou, M. S. (2011). Is banner blindness genuine? Eye tracking internet text advertising. *Applied Cognitive Psychology*, 25(5), 708–716. doi:10.1002/acp.1742
- Honest Ads Act, S. (2017). 1989, 115th Cong. <https://www.congress.gov/bill/116th-congress/senate-bill/1356>
- Hoofnagle, C. J., & Meleshinsky, E. (2015). *Native advertising and endorsement: Schema, source-based misleadingness, and omission of material facts*. TECHNOLOGY SCIENCE. <https://techscience.org/a/2015121503/>
- Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, 47(8), 459–469. doi:10.1111/jasp.12453
- Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2021). Adapting security warnings to counter online disinformation. *Proceedings of the 30th USENIX Security Symposium*, 1163–1180.
- Kim, S. C., Vraga, E. K., & Cook, J. (2021). An eye tracking approach to understanding misinformation and correction strategies on social media: The mediating role of attention and credibility to reduce HPV vaccine misperceptions. *Health Communication*, 36(13), 1687–1696. doi:10.1080/10410236.2020.1787933
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, Supplement*, 13(3), 106–131. doi:10.1177/1529100612451018
- Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, 00(00), 1–38. doi:10.1080/10463283.2021.1876983
- Linville, D. L., & Warren, P. L. (2020). Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, 37(4), 447–467. doi:10.1080/10584609.2020.1718257
- Liu, P. J., Wisdom, J., Roberto, C. A., Liu, L. J., & Ubel, P. A. (2014). Using behavioral economics to design more effective food policies to address obesity. *Applied Economic Perspectives and Policy*, 36(1), 6–24. doi:10.1093/aep/ppt027
- Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2020, October). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments.

- Journal of Experimental Psychology: Applied*, 27(1), 1–16. doi:10.1037/xap0000315
- McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology*, 63(2), 326–332. doi:10.1037/h0048344
- Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (2021). Cognitive reflection correlates with behavior on Twitter. *Nature Communications*, 12(1), 1–10. doi:10.1038/s41467-020-20043-0
- Nieminen, S., & Sankari, V. (2021). Checking politiFact's fact-checks. *Journalism Studies*, 22(3), 358–378. doi:10.1080/1461670X.2021.1873818
- Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, 42(3), 939–960. doi:10.1007/s11109-019-09528-x
- Owens, J., Chaparro, B., & Palmer, E. (2011). Text advertising blindness: The new banner blindness? *Journal of Usability Studies*, 6(3), 172–197.
- Papakyriakopoulos, O., & Goodman, E. P. (2022). The impact of twitter labels on misinformation spread and user engagement: Lessons from Trump's election tweets (February 22, 2022). *Forthcoming in ACM WWW '22*. Available at SSRN. <https://ssrn.com/abstract=4036042>
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*, 66(11), 4944–4957. doi:10.1287/mnsc.2019.3478
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590–595. doi:10.1038/s41586-021-03344-2
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770–780. doi:10.1177/0956797620939054
- Pennycook, G., & Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188(September 2017), 39–50. doi:10.1016/j.cognition.2018.06.011
- Pluviano, S., Della Sala, S., & Watt, C. (2020). The effects of source expertise and trustworthiness on recollection: The case of vaccine misinformation. *Cognitive Processing*, 21(3), 321–330. doi:10.1007/s10339-020-00974-8
- Research, I. O. (2020). *Broad support for raising minimum wage*. <https://www.ioresearch.nl/actueel/grote-steun-voor-verhoging-minimumloon/>
- Richards, A. S., Banas, J. A., & Magid, Y. (2017). More on inoculating against reactance to persuasive health messages: The paradox of threat. *Health Communication*, 32(7), 890–902. doi:10.1080/10410236.2016.1196410
- Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. New York: Farrar, Straus, and Giroux.
- Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of pennycook et al. (2020). *Psychological Science*. doi:10.1177/09567976211024535
- Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of “inoculation” can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, 1(2), 1–23. doi:10.37016/mr-2020-008
- Scrinis, G., & Parker, C. (2016). Front-of-pack food labeling and the politics of nutritional nudges. *Law and Policy*, 38(3), 234–249. doi:10.1111/lapo.12058
- Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning Memory and Cognition*, 43(12), 1948–1961. doi:10.1037/xlm0000422
- Swire-Thompson, B., Ecker, U. K. H., Lewandowsky, S., & Berinsky, A. J. (2020). They might be a liar but they're my liar: Source evaluation and the prevalence of misinformation. *Political Psychology*, 41(1), 21–34. doi:10.1111/pops.12586
- Thorndike, A. N., Riis, J., Sonnenberg, L. M., & Levy, D. E. (2014). Traffic-light labels and choice architecture: Promoting healthy food choices. *American Journal of Preventive Medicine*, 46(2), 143–149. doi:10.1016/j.amepre.2013.10.002
- Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication*, 33(3), 460–480. doi:10.1080/10584609.2015.1102187
- Traberg, C. S., & van der Linden, S. (2022). Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. *Personality and Individual Differences*, 185 (September 2021), 111269. doi:10.1016/j.paid.2021.111269
- Tucker, J. A., Guess, A., Barbera, P., Vaccari, C., Siegel, A. . . . Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. *Hewlett Foundation*, (Issue March), 1–95. doi:10.2139/ssrn.3144139
- van Drunen, M. Z., Groen-Reijman, E., Dobber, T., Noroozian, A., Leerssen, P. J. . . . Votta, F. A. (2022). Transparency and (no) more in the political advertising regulation. *Internet Policy Review*. doi:10.14763/2022.1.1652
- Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society*, 22(1), 29–48. doi:10.1080/15205436.2018.1511807
- van Reijmersdal, E. A., Rozendaal, E., Hudders, L., Vanwesenbeeck, I., Cauberghe, V., & van Berlo, Z. M. C. (2020). Effects of disclosing influencer marketing in videos: An eye tracking study among children in early adolescence.

- Journal of Interactive Marketing*, 49(1), 94–106. doi:10.1016/j.intmar.2019.09.001
- Vraga, E. K., Bode, L., & Tully, M. (2020). Creating news literacy messages to enhance expert corrections of misinformation on twitter. *Communication Research*, 49(2), 245–267. doi:10.1177/0093650219898094
- Vraga, E. K., Kim, S. C., Cook, J., & Bode, L. (2020). Testing the effectiveness of correction placement and type on instagram. *The International Journal of Press/politics*, 25(4), 632–652. doi:10.1177/1940161220919082
- Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3), 350–375. doi:10.1080/10584609.2019.1668894
- Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2), 155–177. doi:10.1177/0093650219854600
- Wojdyski, B. W., & Evans, N. J. (2016). Going native: Effects of disclosure position and language on the recognition and evaluation of online native advertising. *Journal of Advertising*, 45(2), 157–168. doi:10.1080/00913367.2015.1115380
- Zerback, T., Töpfl, F., & Knöpfle, M. (2020). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media and Society*, 23(5), 1080–1098. doi:10.1177/1461444820908530
- Zuwerink Jacks, J., & Cameron, K. A. (2003). Strategies for resisting persuasion. *Basic and Applied Social Psychology*, 25(2), 145–161. doi:10.1207/s15324834basps250