Displaying News Source Trustworthiness Ratings Reduces Sharing Intentions for False News Posts

Tatiana Celadin, Valerio Capraro, Gordon Pennycook, David G. Rand

Abstract. Professional fact-checking of individual news headlines is an effective way to fight misinformation, but it is not easily scalable, because it cannot keep pace with the massive speed at which news content gets posted on social media. Here we provide evidence for the effectiveness of ratings of news sources, instead of individual news articles. In a large pre-registered experiment with quota-sampled Americans, we find that participants are less likely to share false headlines (and are more discerning of true versus false headlines) when 1-to-5 star source trustworthiness ratings were applied to news headlines. This is true when the ratings are generated both by fact-checkers and by laypeople (although the effect is stronger using fact-checker ratings). We also observe a positive spillover effect: sharing discernment also increases for headlines whose source was not rated, likely because the presence of ratings on some headlines prompts users to reflect on source quality more generally. This study suggests that displaying information regarding the trustworthiness of news sources provides a scalable approach for reducing the spread of low-quality information.

1 Introduction

While false news represents only a small proportion of news that is consumed and shared on social media every day (Guess, Nagler, and Tucker 2019; Allen et al. 2020), there is substantial concern about the consequences of exposure to even small amounts of misinformation. For example, exposure can increase beliefs (Fazio et al. 2015), even for extremely implausible political claims (Pennycook, Cannon, and Rand 2018), and COVID-19 misinformation can reduce vaccination intentions (Loomba et al. 2021). As a result, reducing false-news sharing on social media is considered by many experts to be one of the most pressing contemporary issues (Kuklinski et al. 2000; Lewandowsky et al. 2012; Lazer et al. 2018; Pennycook and Rand 2021).

One promising approach to combatting misinformation is professional fact-checking,

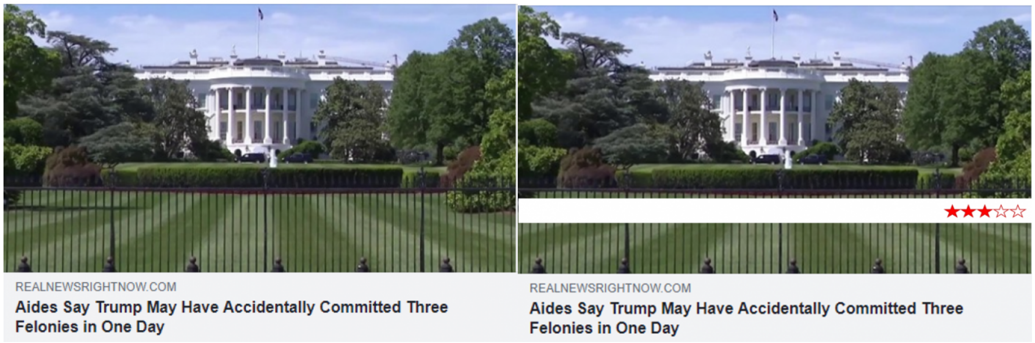
doi:10.54501/jots.v1i5.100

provided by specialized “fact-checkers,” who check the veracity of specific pieces of content that are posted online, find false content, and label it with suitable warnings. Despite initial concerns that corrections may backfire (Nyhan and Reifler 2010), there is now increasing evidence that tagging inaccurate news with warnings tends to increase people’s ability to discern truth from falsehood (Pennycook, Cannon, and Rand 2018; Wood and Porter 2019; Pennycook, McPhetres, et al. 2020; Swire-Thompson et al. 2020; Yaqub et al. 2020). Additionally, the effect of fact-checkers can be reinforced by social media platforms that can use ranking algorithms to reduce the likelihood of exposing users to content labeled as false by fact-checkers. However, despite these promises, fact-checking suffers from a major limitation: scalability.

Professional fact-checking of news headlines requires considerable time and effort from a small number of experts. Thus, it is difficult for professional fact-checking to be implemented at a scale that matches the massive rate at which news content gets posted online. The consequent lack of coverage and delay between publication of a news headline and its fact-checking have several negative consequences. One is that false news often gets labeled as false after it has already been read by thousands, if not millions, of users. Another is that the contemporary presence of tagged and untagged news may lead users to assume that news that is not marked as false is actually true; this “implied truth effect” might lead to an increase in false-news belief and sharing (Pennycook, Bear, et al. 2020). Additionally, the fact that social media users tend to be connected to like-minded others (Colleoni, Rozza, and Arvidsson 2014; Flaxman, Goel, and Rao 2016; Brady et al. 2017; Stewart et al. 2019; Mosleh et al. 2021) implies that corrections often reach a different audience from the one reached by the original, untagged, headline (Guess et al. 2020). On top of that, fact-checkers’ evaluations are not trusted by a substantial proportion of people (Flamini 2019), which many scholars worry may limit the effectiveness of fact-checks.

A possible solution that has been proposed to scale-up fact-checking is by harnessing the “wisdom of crowds.” This method is based on classic work showing that aggregating judgments from the crowd outperforms experts’ judgments in a wide variety of domains (Galton 1907; Surowiecki 2005). In the context of news headline ratings, recent work found that accuracy ratings provided by laypeople correlate strongly with accuracy ratings provided by fact-checkers (Allen et al. 2021; Resnick et al. 2021). Although other work was less optimistic about the ability of crowds to identify misinformation (Godel et al. 2021), these differing conclusions appear to arise largely from different analytic approaches; when using crowds to generate scalar (rather than binary) classifications, the data from both Allen et al. (2021) and Godel et al. (2021) indicate that small crowds can perform as well or better than a professional fact-checker (Martel et al. 2022). This suggests that a potential way to scale-up fact-checking is to collect ratings of article accuracy from laypeople instead of fact-checkers. Yet, given the vast amount of content posted on social media every moment, even this approach may have difficulty in fully meeting the scale of the misinformation identification challenge online. Indeed, as for fact-checking, laypeople ratings will also experience a systematic delay relative to the time when a news story gets published.

Here we study an alternative approach to reducing the spread of misinformation that aims to be more scalable by focusing on ratings of news *sources*, rather than individual news articles. Specifically, we study whether displaying 1-to-5 star trustworthiness ratings of news sources can decrease the sharing of false content. This intervention would be more easily scalable than the corresponding intervention based on accuracy ratings at the headline level, because sources of news are far fewer (and are created at a much slower pace) than news headlines. See Figure 1 for a sample method. We opted for this intervention because there is evidence that these ratings have a direct influence on individuals' behavior in consumer contexts (Chevalier and Mayzlin 2006; Luca and Zervas 2016), and thus we hypothesized that this rating system could be very effective when users face news headlines on social media. Of course, this approach would also have the downside of being coarse, as low-credibility sources sometimes publish accurate news and high-credibility sources sometimes publish inaccurate news (e.g., Dias, Pennycook, and Rand 2020; Stewart et al. 2021).



REALNEWSRIGHTNOW.COM

Aides Say Trump May Have Accidentally Committed Three Felonies in One Day

REALNEWSRIGHTNOW.COM

Aides Say Trump May Have Accidentally Committed Three Felonies in One Day

(a)

(b)

Figure 1: This figure displays how news headlines were presented. (a) Sample headline in the Baseline condition. (b) Sample headline in the Laypeople condition; headlines in the Fact-checkers condition were displayed in the same fashion as in the Laypeople condition.

Related to our work, Aslett et al. (2022) found that displaying small categorical credibility labels (a green shield for reliable sources, a red shield for unreliable sources, a gray shield for user-generated content, and a golden shield for satire) had no overall effect on *consumption* of news from reliable versus unreliable sources, but it may have reduced consumption of unreliable sources among users who consumed a large amount of unreliable news at baseline. Here, we examine news *sharing* intentions rather than consumption. Reducing sharing is particularly important given that one person sharing a piece of news potentially exposes a substantial number of others to that news. And sharing has been found to be a generally malleable outcome (perhaps more so than consumption), as sharing decisions are often made without careful prior consideration (Pennycook et al. 2021; Pennycook and Rand 2021; Arechar et al. 2022; Capraro and Celadin 2022). Furthermore, we test a different, more detailed (and potentially more salient) form of labeling technique inspired by popular online rating systems, such as TripAdvisor and Google. Kim, Moravec, and Dennis (2019) tested a similar labeling technique on perceived believability of news about abortion created by the authors. They found that low ratings have stronger effects than high ratings and the effect is even stronger when the ratings are provided by fact-checkers rather than laypeople. However, Kim, Moravec, and Dennis (2019) could not analyze the effect of the labels on the ability to tell truth from falsehoods (as all their headlines were invented). This is a crucial limitation given the practical importance of testing the differential effect of the intervention on true versus false headlines. Moreover, they only presented headlines on one specific topic (abortion), which poses questions about the generalizability of their results. In our study we overcome these limitations by using true and false headlines on a range of topics that were actually posted online (rather than fabricated by the researchers).

As for the source of ratings, we contrast the effectiveness of source ratings provided by fact-checkers versus layperson crowds. Although prior work has shown expert and crowd ratings of news-source credibility to be very highly correlated (Pennycook and Rand 2019; Epstein, Pennycook, and Rand 2020), it is important to compare the effect of displaying these ratings to shed light on which ratings, if any, have a bigger effect, given concerns that fact-checkers may not be trusted by a substantial proportion of people (Flamini 2019), but also that experts may be seen as more legitimate than crowds for content moderation (Pan et al. 2022). We also examine the effect that placing the trustworthiness rating on some headlines (direct effect) has on headlines that do not have ratings (indirect effect). Specifically, in the treatment conditions, half of the headlines are presented with ratings while half of the headlines are presented without ratings. From a theoretical perspective, this allows us to differentiate the impact of the actual information about the source quality (which is only present for the labeled headlines) from more generally priming users to think about accuracy (Pennycook et al. 2021), which will be the case for all headlines in the treatments. From a practical perspective, new or niche content producers may be unfairly penalized by crowd ratings (as people typically distrust unfamiliar sources; Pennycook and Rand (2019)), and thus it may be desirable to present relatively unknown sources without ratings. This design allows us to assess the effect on those unlabeled sites.

2 Method

We collected N = 2,250 US-based participants from Lucid (quota-sampled to match the national distribution on age, gender, ethnicity, and geographic region). After providing their informed consent, participants faced two attention checks. As pre-registered, we eliminated subjects who fail one or both attention checks, after which 1,627 participants remained. Participants were then asked two questions about social media use. Specifically, they were asked which types of content, if any, they consider sharing on social media (available answers: political news, sport news, celebrity news, science/technology news, business news, other) and what type of social media accounts they use (available answers: Facebook, Instagram, Twitter, WhatsApp, Snapchat, other). Next, participants were randomly assigned to one of three conditions: the *Baseline*, the *Fact-Checkers*, and the *Laypeople*.

In the *Baseline* condition, participants were shown 24 news headlines in Facebook format, 12 true and 12 false, in random order. We collected 12 pro-democrat (6 true and 6 false) and 12 pro-republican (6 true and 6 false) news headlines. False headlines were collected from seven news sources (clashdaily.com, downtrend.com, conservativedailypost.com, realnewsrightnow.com, now8news.com, bb4sp.com, yournewswire.com) which have been classified by Zimdars (2016) as "sources that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports." True headlines were collected from sources that have been rated as highly reliable by professional fact-checkers. For each news headline, subjects were asked: "If you were to see the above on social media, how likely would you be to share it?" (Available answers: 1 = "extremely unlikely," 2 = "moderately unlikely," 3 = "slightly unlikely," 4 = "slightly likely" 5 = "moderately likely," 6 = "extremely likely.") In the analysis, the answers are renormalized between 0 and 1 for ease of interpretation of the coefficients.

The *Fact-Checkers* condition was similar to the Baseline, with the following two differences: (1) prior to seeing the news headlines, subjects were informed that "For some news headlines, you will see a trustworthiness rating of the source of the headline. The rating system goes from 1 to 5 stars, where 1 star means that the news outlet is very unreliable while 5 stars means that the outlet is very reliable. The ratings were created by a set of eight professional fact-checkers." These ratings were collected by Pennycook and Rand (2019), who asked eight professional fact-checkers to rate the sources; (2) for 12 of these 24 news headlines, we added a banner reporting the number of stars received by the corresponding news outlet. Ratings were generated from Pennycook and Rand (2019). News headlines that were shown with ratings were counterbalanced across subjects and were shown in random order. See Figure 1 for a sample method.

The *Laypeople* condition was similar to the Fact-Checker condition, with the difference that the information about the trustworthiness ratings was given through the following sentence: "For some news headlines, you will see a trustworthiness rating of the source of the headline. The rating system goes from 1 to 5 stars, where 1 star means that the news outlet is very unreliable while 5 stars means that the outlet is very reliable. The ratings were created by a set of 970 laypeople." These ratings were also collected by Pennycook and Rand (2019).

After reporting their sharing intentions, participants entered a demographic questionnaire, where they were asked a number of standard questions, including "which of the following best describes your political preferences?" (Available answers: Strongly Democratic, Democratic, Lean Democratic, Lean Republican, Republican, Strongly Republican).

Pre-registration,[1] screenshots of all news headlines, data, and analysis code are available online (https://osf.io/9muw7/). This study was deemed exempt by the MIT Committee on the Use of Humans as Experimental Subjects, protocol E-4195.

3 Results

The average willingness to share across conditions (Baseline vs. Laypeople vs. Fact-checkers) and headline veracity (False vs. True) is shown in Figure 2.

We begin by comparing sharing intentions for headlines with ratings in the treatments versus headlines in the control. This allows us to assess the direct impact of the ratings. We find that sharing discernment (the difference in sharing willingness of true headlines relative to false headlines) was significantly increased by both the fact-checker ratings (b = 0.090, t = 7.56, 95% CI = [0.067, 0.113], p < 0.001) and the layperson ratings (b = 0.052, t = 4.64, 95% CI = [0.030, 0.075], p < 0.001), as seen in Figure 3a. In particular, the sharing of false headlines was significantly reduced by both the fact-checker ratings (b = -0.084, t = -4.91, 95% CI = [-0.118, -0.050], p < 0.001) and the layperson ratings (b = -0.042, t = -2.39, 95% CI = [-0.076, -0.008], p = 0.017), while the sharing of true headlines was not affected by either the fact-checker ratings (b = 0.006, t = 0.33, 95% CI = [-0.029, 0.041], p = 0.743) or the layperson ratings (b = 0.011, t = 0.61, 95% CI = [-0.024, 0.046], p = 0.542), as seen in Figure 3b. Importantly, for both sharing discernment and the sharing of false headlines, the rating effect size was significantly larger for the fact-checker ratings than for the layperson ratings (sharing discernment: F(1, 1626) = 7.41, p = 0.007; sharing of false headlines: F(1, 1626) = 6.38, p = 0.012), while there was no difference for the sharing of true headlines (F(1, 1626) = 0.08, p = 0.782). Thus, we find that source-level ratings may help improve the quality of news shared, with a particular advantage of fact-checker ratings.

We then turn to comparing sharing intentions for headlines without ratings in the

[1] We deviate from the pre-registration in two ways: (1) we add two regressors: Lp Unrated and Fc Unrated. These regressors allow us to study the indirect effect of the ratings on the willingness to share unrated headlines. It was indeed our intention to study this indirect effect (otherwise we would have not placed the ratings on only half of the headlines), but we forgot to pre-register it. (2) we changed the econometric model, in order to align it with the most recent research on the topic (e.g., Pennycook et al. 2021). Specifically, instead of using mixed-effects linear regression, we use linear regression with clustered standard errors. Due to too few clusters, we cluster only at the participant level and not at the headline level. We note that we obtain qualitatively similar results if, instead of using linear regression, we use a mixed-effects regression.

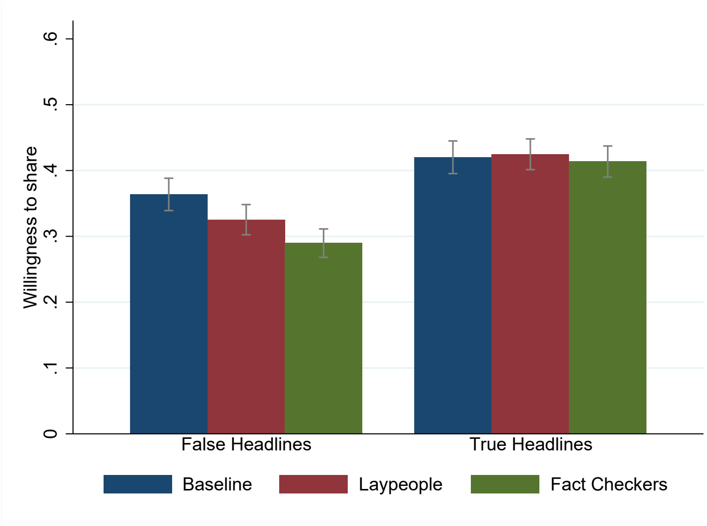
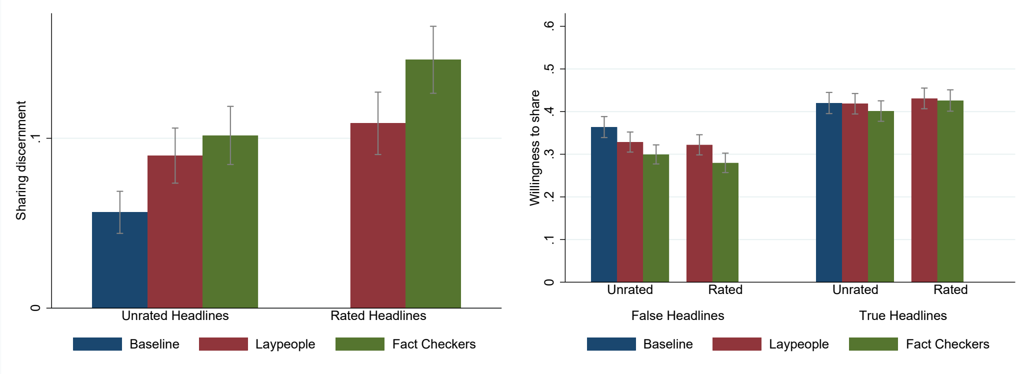


Figure 2: Willingness to share (normalized between 0 and 1) split by headline veracity (False vs. True) and condition (Baseline vs. Laypeople vs. Fact-Checkers). Error bars represent 95% CI clustered at the participant level.



(a)

(b)

Figure 3: (a) Sharing discernment split by condition (Baseline vs. Laypeople vs. Fact-Checkers), and presence of the trustworthiness rating (Rated vs. Unrated). (b) Willingness to share (normalized between 0 and 1) split by headline veracity (False vs. True) and condition (Baseline vs. Laypeople vs. Fact-Checkers), and by presence of the trustworthiness rating (Rated vs. Unrated). Error bars represent 95% CI clustered at the participant level.

treatment to headlines in the control. This allows us to assess the indirect spillover impact of having ratings present on other news headlines (without providing actual content-specific information for the headlines in question). The results are qualitatively similar to the direct rating effects: Sharing discernment was significantly higher for unrated headlines in the treatment compared to headlines in the control in both the fact-checker condition (b = 0.045, t = 4.20, 95% CI = [0.024, 0.066], p < 0.001) and the layperson condition (b = 0.033, t = 3.21, 95% CI = [0.013, 0.054], p = 0.001), as seen in Figure 3a. Specifically, sharing of false headlines was significantly lower for unrated headlines in the treatment compared to control in the fact-checker condition (b = -0.064, t = -3.78, 95% CI = [-0.097, -0.031], p < 0.001) and the layperson condition (b = -0.035, t = -2.02, 95% CI = [-0.069, -0.001], p = 0.043), whereas there was no effect in the sharing of true headlines for unrated headlines in both treatments (b = -0.019, t = -1.07, 95% CI = [-0.053, 0.016], p = 0.284; b = -0.002, t = -0.10, 95% CI = [-0.036, 0.033], p = 0.922), as seen in Figure 3b. The magnitude of these indirect effects was not significantly different between the fact-checker and layperson ratings (sharing discernment: F(1, 1626) = 0.98, p = 0.323; sharing of false headlines: F(1, 1626) = 3.08, p = 0.080; sharing of real news: F(1, 1626) = 0.98 p = 0.322). Thus, we find that applying source-level ratings to half of the headlines improves the quality of information sharing even for headlines whose sources were not rated. The lack of difference between the fact-checkers and laypeople for the indirect effects is consistent with the interpretation that the positive spillover effect is occurring due to a general priming of the concept of accuracy, rather than the specific information provided.

Finally, we compare the magnitude of the direct and indirect effects. Importantly, we find that the direct effects of the ratings are significantly larger than the indirect effects (fact-checker: F(1, 1626) = 21.46, p < 0.001; layperson: F(1, 1626) = 5.00, p = 0.026). This demonstrates that the ratings are not only priming users to consider accuracy—the actual labels' content also helps to improve sharing discernment by informing users about source credibility.

In the Appendix, we demonstrate that these results are qualitatively similar if we exclude participants who failed a post-experimental attention check or exclude participants who indicated that they never share political news online. We also examine whether the rating effects are moderated by political concordance of the news, or political partisanship. The only significant moderation we observe is that the fact-checker ratings increase sharing discernment significantly more for politically concordant headlines compared to politically discordant ones. This is encouraging, in so much as people are more likely to share politically concordant news than politically discordant ones (Faragó, Kende, and Krekó 2019; Pennycook and Rand 2019; Vegetti and Mancosu 2020; Pennycook et al. 2021). Political partisanship was asked at the end of the survey. However, there are concerns about asking this kind of question at the end of the survey (Montgomery, Nyhan, and Torres 2018); thus, we checked whether the political partisanship is balanced across treatments. Overall, there was no difference between treatments (Wilcoxon rank-sum test: control vs laypeople=0.635; control vs fact-checkers: p=0.753; fact-checkers vs laypeople p=0.888). Finally, we also check the duration of the effect of the intervention (Roozenbeek, Freeman, and Linden 2021; Pennycook and Rand 2022; Capraro and Celadin 2022) by using the headlines display order as a proxy of time; overall, we found no consistent evidence of an order effect, suggesting that the treatment effects are relatively stable at least for the duration of the study (see Table 3 and Figure 4 in the Appendix).

4 Discussion

In this study, we have shown that labeling news headlines with a source trustworthiness rating can help reduce the spread of misinformation. Specifically, we found that the trustworthiness ratings of news sources increase sharing discernment and decrease false-headlines sharing, compared to the baseline, both when they are provided by fact-checkers and when they are provided by laypeople. The effect of these labels is larger when they are provided by fact-checkers compared to when they are provided by laypeople. Additionally, we found similar results when we considered the indirect, spillover effect of the trustworthiness ratings: they tend to also improve sharing discernment among news that are presented without any rating. The magnitude of the indirect effect of ratings provided by fact-checkers is close to the one of ratings provided by laypeople. Finally, the direct effects are significantly larger than the indirect effects.

These results suggest that adding a news source trustworthiness rating on news headlines can decrease the sharing intention of false news and increase sharing discernment, especially when the ratings come from fact-checkers. This intervention has two main upsides relative to other approaches. First, it is scalable, because it is based on the *source* of information, instead of each single news headline. Sources of information are much fewer than news headlines, and new sources of information are created at a much slower rate than news headlines. Thus, this intervention can more easily keep pace with the launch of new sources of information than attempts to assess individual pieces of content. The second upside is that this intervention does not seem to generate a perverse “implied truth effect” for unlabeled sources. On the contrary, the pattern is the opposite of implied truth: the intervention has a similar positive effect on both rated and unrated news (although smaller in magnitude for the unrated news). This is an important point, because new or niche sources of content might be penalized from our rating procedure (especially if provided by laypeople), because laypeople tend to distrust unknown sources of information (Pennycook and Rand 2019). Therefore, it might be desirable to present new or niche sources without rating. Our results show that this is not an issue, as the main effects spill over to news without ratings.

In addition to the related work on labeling source quality reviewed in the introduction (Kim, Moravec, and Dennis 2019; Aslett et al. 2022), other work has looked at the effect of highlighting the source of news (Jakesch et al. 2018; Dias, Pennycook, and Rand 2020; Pennycook and Rand 2020) and found no effect. Typically, these works highlight only the source of information by, for example, displaying it with larger characters or a logo (Dias, Pennycook, and Rand 2020), but without providing explicit information regarding whether it is trustworthy or not. Our study differs from these because we provide direct information about source trustworthiness, which appears to have much more effect than simply emphasizing the source. Our work also connects to previous research investigating how trustworthy fact-checkers are perceived to be. In this regard, it was suggested that some people distrust fact-checkers, implying that fact-checker interventions may be less effective than interventions based on laypeople (Flamini 2019). Yet other work found that, when labeling individual articles, fact-checker ratings were more effective than crowd ratings (Kim, Moravec, and Dennis 2019; Yaqub et al. 2020), and that experts are seen as more legitimate than crowds for deciding what content to moderate (Pan et al. 2022). Our results are in line with the latter findings, as both our interventions worked, but the one based on fact-checkers had a stronger effect. More generally, our study connects to the literature on accuracy salience and provides yet another confirmation that accuracy salience improves sharing discernment (Pennycook, McPhetres, et al. 2020; Byles et al. 2021; Epstein et al. 2021; Pennycook et al. 2021; Roozenbeek, Freeman, and Linden 2021; Capraro and Celadin 2022;

Rasmussen, Lindekilde, and Petersen 2022). See Pennycook and Rand (2022) for a meta-analysis.

Our study suffers from some limitations. The first is related to the ecological validity: our intervention was tested in a context that does not perfectly reproduce any social media platform; moreover, we studied hypothetical sharing intentions rather than actual sharing. Previous work suggests that sharing intentions collected on survey experiments tend to be correlated with actual sharing decisions and show similar patterns of correlation with headline features (Mosleh, Pennycook, and Rand 2020), and that accuracy-based interventions tend to work in a similar fashion in survey experiments and social media field experiments (Pennycook et al. 2021). Future work should investigate the impact of source label credibility on actual sharing on social media. The second limitation is related to generalizability. We recruited a quota-matched sample of people living in the USA to evaluate a specific set of headlines. Future work should assess generalizability to other headlines and topics, and to other countries (Arechar et al. 2022). Moreover, it would be of interest to investigate whether trustworthiness ratings affect other dimensions, 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 (Aslett et al. 2022).

Two more limitations are related to the source trustworthiness rating per se. First, low-quality sources can sometimes publish true news headlines, but users may distrust and not share these news headlines because they come from non-credible sources; conversely, users may believe and share false news published by credible sources. Second, there is evidence that these ratings can be altered by interested parties to damage their competitor or advantage themselves. Such behavior could undermine the potential of trustworthiness rating (Smith 2004; Segal 2011; Luca and Zervas 2016). These limitations highlight a point that has been made several times (Pennycook and Rand 2021; Bak-Coleman et al. 2022): that it is unlikely that a single intervention alone can solve the problem of misinformation. In particular, this source-based trustworthiness rating, if implemented, should be accompanied with other interventions.

Despite these limitations, our results suggest that adding source trustworthiness ratings to news posts may increase sharing discernment, specifically by decreasing the intention to share false headlines. Therefore, this appears to be a promising approach for social media platforms to consider in their efforts to reduce the spread of misinformation at scale.

References

Allen, Jennifer, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. “Scaling up fact-checking using the wisdom of crowds.” *Science Advances* 7 (36): eabf4393. https://doi.org/https://doi.org/10.2139/ssrn.3502581.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2020. “Evaluating the fake news problem at the scale of the information ecosystem.” *Science Advances* 6 (14): eaay3539.

Arechar, Antonio Alonso, Jennifer Nancy Lee Allen, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael Stagnaro, Jerry Zhang, et al. 2022. “Understanding and reducing online misinformation across 16 countries on six continents.” *PsyArXiv,* https://doi.org/https://doi.org/10.31234/osf.io/a9frz.

Aslett, Kevin, Andrew M Guess, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2022. “News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions.” *Science Advances* 8 (18): eabl3844. https://doi.org/https://doi.org/10.1126/sciadv.abl3844.

Bak-Coleman, Joseph B, Ian Kennedy, Morgan Wack, Andrew Beers, Joseph S Schafer, Emma S Spiro, Kate Starbird, and Jevin D West. 2022. “Combining interventions to reduce the spread of viral misinformation.” *Nature Human Behaviour* 6 (10): 1372–80. https://doi.org/https://doi.org/10.1038/s41562-022-01388-6.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. “Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys.” *American journal of political science* 58 (3): 739–53. https://doi.org/https://doi.org/10.1111/ajps.12081.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. “Emotion shapes the diffusion of moralized content in social networks.” *Proceedings of the National Academy of Sciences* 114 (28): 7313–18. https://doi.org/https://doi.org/10.1073/pnas.1618923114.

Byles, Oliver, Joshua Calianos, Sade Francis, Chun Hey Brian Kot, H Nephi Seo, and Brendan Nyhan. 2021. “The Effects of Accuracy Salience and Affective Polarization on Truth Discernment in Online News Sharing.” *Unpublished manuscript.*

Capraro, Valerio, and Tatiana Celadin. 2022. “‘I think this news is accurate’: Endorsing accuracy decreases the sharing of fake news and increases the sharing of real news.” *Personality and Social Psychology Bulletin,* 01461672221117691. https://doi.org/https://doi.org/10.31234/osf.io/s3q5n.

Chevalier, Judith A, and Dina Mayzlin. 2006. “The effect of word of mouth on sales: Online book reviews.” *Journal of Marketing Research* 43 (3): 345–54. https://doi.org/https://doi.org/10.1509/jmkr.43.3.345.

Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson. 2014. “Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data.” *Journal of Communication* 64 (2): 317–32. https://doi.org/https://doi.org/10.1111/jcom.12084.

Dias, Nicholas, Gordon Pennycook, and David G Rand. 2020. “Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media.” *Misinformation Review,* https://doi.org/https://doi.org/10.37016/mr-2020-001.

Epstein, Ziv, Adam J Berinsky, Rocky Cole, Andrew Gully, Gordon Pennycook, and David G Rand. 2021. “Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online.” *Misinformation Review,* https://doi.org/https://doi.org/10.31234/osf.io/sjfbn.

Epstein, Ziv, Gordon Pennycook, and David Rand. 2020. “Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources.” In *Proceedings of the 2020 CHI conference on human factors in computing systems,* 1–11. https://doi.org/https://doi.org/10.1145/3313831.3376232.

Faragó, Laura, Anna Kende, and Péter Krekó. 2019. “We only believe in news that we doctored ourselves.” *Social Psychology,* https://doi.org/https://doi.org/10.1027/1864-9335/a000391.

Fazio, Lisa K, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh. 2015. “Knowledge does not protect against illusory truth.” *Journal of Experimental Psychology: General* 144 (5): 993. https://doi.org/https://doi.org/10.1037/xge0000098.

Flamini, Daniela. 2019. “Most Republicans don’t trust fact-checkers, and most Americans don’t trust the media.” *Poynter.*

Flaxman, Seth, Sharad Goel, and Justin M Rao. 2016. “Filter bubbles, echo chambers, and online news consumption.” *Public Opinion Quarterly* 80 (S1): 298–320. https://doi.org/https://doi.org/10.1093/poq/nfw006.

Galton, Francis. 1907. “Vox Populi.” *Nature* 75:450–51. https://doi.org/https://doi.org/10.1038/075450a0.

Godel, William, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. “Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking.” *Journal of Online Trust and Safety* 1 (1). https://doi.org/https://doi.org/10.54501/jots.v1i1.15.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. “Less than you think: Prevalence and predictors of fake news dissemination on Facebook.” *Science advances* 5 (1): eaau4586. https://doi.org/https://doi.org/10.1126/sciadv.aau4586.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. “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): 15536–45. https://doi.org/https://doi.org/10.1073/pnas.1920498117.

Jakesch, Maurice, Moran Koren, Anna Evtushenko, and Mor Naaman. 2018. “The role of source, headline and expressive responding in political news evaluation.” *Headline and Expressive Responding in Political News Evaluation,* https://doi.org/https://doi.org/10.2139/ssrn.3306403.

Kim, Antino, Patricia L Moravec, and Alan R Dennis. 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): 931–68. https://doi.org/ttps://doi.org/10.1080/07421222.2019.1628921.

Kuklinski, James H, Paul J Quirk, Jennifer Jerit, David Schwieder, and Robert F Rich. 2000. “Misinformation and the currency of democratic citizenship.” *The Journal of Politics* 62 (3): 790–816. https://doi.org/https://doi.org/10.1111/0022-3816.00033.

Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. "The science of fake news." *Science* 359 (6380): 1094–96. https://doi.org/https://doi.org/10.1126/science.aao2998.

Lewandowsky, Stephan, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. "Misinformation and its correction: Continued influence and successful debiasing." *Psychological science in the public interest* 13 (3): 106–31. https://doi.org/https://doi.org/10.1177/1529100612451018.

Loomba, Sahil, Alexandre de Figueiredo, Simon J Piatek, Kristen de Graaf, and Heidi J Larson. 2021. "Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA." *Nature human behaviour* 5 (3): 337–48. https://doi.org/https://doi.org/10.1038/s41562-021-01056-1.

Luca, Michael, and Georgios Zervas. 2016. "Fake it till you make it: Reputation, competition, and Yelp review fraud." *Management Science* 62 (12): 3412–27.

Martel, Cameron, Jennifer Nancy Lee Allen, Gordon Pennycook, and David Rand. 2022. "Crowds can effectively identify misinformation at scale."

Montgomery, Jacob M, Brendan Nyhan, and Michelle Torres. 2018. "How conditioning on posttreatment variables can ruin your experiment and what to do about it." *American Journal of Political Science* 62 (3): 760–75. https://doi.org/https://doi.org/10.1111/ajps.12357.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G Rand. 2021. "Shared partisanship dramatically increases social tie formation in a Twitter field experiment." *Proceedings of the National Academy of Sciences* 118 (7): e2022761118. https://doi.org/https://doi.org/10.1073/pnas.2022761118.

Mosleh, Mohsen, Gordon Pennycook, and David G Rand. 2020. "Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter." *PLOS one* 15 (2): e0228882. https://doi.org/https://doi.org/10.1371/journal.pone.0228882.

Nyhan, Brendan, and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32 (2): 303–30. https://doi.org/https://doi.org/10.1007/s11109-010-9112-2.

Pan, Christina A, Sahil Yakhmi, Tara P Iyer, Evan Strasnick, Amy X Zhang, and Michael S Bernstein. 2022. "Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries." *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW1): 1–31. https://doi.org/https://doi.org/10.1145/3512929.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 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): 4944–57. https://doi.org/https://doi.org/10.1287/mnsc.2019.3478.

Pennycook, Gordon, Tyrone D Cannon, and David G Rand. 2018. "Prior exposure increases perceived accuracy of fake news." *Journal of experimental psychology: general* 147 (12): 1865. https://doi.org/https://doi.org/10.31234/osf.io/9qdza.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. "Shifting attention to accuracy can reduce misinformation online." *Nature* 592 (7855): 590–95. https://doi.org/https://doi.org/10.1038/s41586-021-03344-2.

Pennycook, Gordon, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. “Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention.” *Psychological science* 31 (7): 770–80. https://doi.org/https://doi.org/10.1177/0956797620939054.

Pennycook, Gordon, and David G Rand. 2019. “Fighting misinformation on social media using crowdsourced judgments of news source quality.” *Proceedings of the National Academy of Sciences* 116 (7): 2521–26. https : //doi.org/https : //doi.org/10.1073/pnas.1806781116.

———. 2020. “Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking.” *Journal of personality* 88 (2): 185–200. https://doi.org/https://doi.org/10.1111/jopy.12476.

———. 2021. “The psychology of fake news.” *Trends in cognitive sciences* 25 (5): 388–402. https://doi.org/https://doi.org/10.1016/j.tics.2021.02.007.

———. 2022. “Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation.” *Nature communications* 13 (1): 2333. https://doi.org/https://doi.org/10.1038/s41467-022-30073-5.

Rasmussen, Jesper, Lasse Lindekilde, and Michael Bang Petersen. 2022. “Public health communication decreases false headline sharing by boosting self-efficacy,” https://doi.org/https://doi.org/10.31234/osf.io/8wdfp.

Resnick, Paul, Aljohara Alfayez, Jane Im, and Eric Gilbert. 2021. “Informed crowds can effectively identify misinformation.” *arXiv preprint arXiv:2108.07898,* https://doi.org/https://doi.org/10.1287/mnsc.2015.2304.

Roozenbeek, Jon, Alexandra LJ Freeman, and Sander van der Linden. 2021. “How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al.(2020).” *Psychological science* 32 (7): 1169–78. https://doi.org/https://doi.org/10.1177/09567976211024535.

Segal, David. 2011. “A rave, a pan, or just a fake?” *The New York Times.*

Smith, David. 2004. “Amazon reviewers brought to book.” *The Guardian,* no. February 14.

Stewart, Alexander J, Antonio A Arechar, David G Rand, and Joshua B Plotkin. 2021. “The distorting effects of producer strategies: Why engagement does not reliably reveal consumer preferences for misinformation.” *arXiv preprint arXiv:2108.13687.*

Stewart, Alexander J, Mohsen Mosleh, Marina Diakonova, Antonio A Arechar, David G Rand, and Joshua B Plotkin. 2019. “Information gerrymandering and undemocratic decisions.” *Nature* 573 (7772): 117–21. https://doi.org/https://doi.org/10.1038/s41586-019-1507-6.

Surowiecki, James. 2005. *The wisdom of crowds.* Anchor.

Swire-Thompson, Briony, Ullrich KH Ecker, Stephan Lewandowsky, and Adam J Berinsky. 2020. “They might be a liar but they’re my liar: Source evaluation and the prevalence of misinformation.” *Political psychology* 41 (1): 21–34. https://doi.org/https://doi.org/10.1111/pops.12586.

Vegetti, Federico, and Moreno Mancosu. 2020. “The impact of political sophistication and motivated reasoning on misinformation.” *Political Communication* 37 (5): 678–95. https://doi.org/https://doi.org/10.1080/10584609.2020.1744778.

Wood, Thomas, and Ethan Porter. 2019. “The elusive backfire effect: Mass attitudes’ steadfast factual adherence.” *Political Behavior* 41:135–63. https://doi.org/https://doi.org/10.1007/s11109-018-9443-y.

Yaqub, Waheeb, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. “Effects of credibility indicators on social media news sharing intent.” In *Proceedings of the 2020 chi conference on human factors in computing systems,* 1–14. https://doi.org/https://doi.org/10.1145/3313831.3376213.

Zimdars, Melissa. 2016. “My ‘Fake News List’ Went Viral. But Made-up Stories are Only Part of the Problem.” *Washington Post,* https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem/.

Authors

Tatiana Celadin is Assistant Professor at Ca’Foscary University.

(tatiana.celadin@unive.it)

Valerio Capraro is Associate Professor at University Milano Bicocca.

(valerio.capraro@unimib.it)

Gordon Pennycook is Associate Professor at University of Regina.

(gordon.pennycook@uregina.ca)

David G. Rand is Full Professor at MIT.

(drand@mit.edu)

Acknowledgements

The authors declare no conflict of interests.

Data availability statement

Replication files are available at: https://osf.io/9muw7/.

The study pre-analysis plan is registered with AsPredicted.

Funding statement

This study was supported by the Middlesex University Business School grant (RK-EFF).

Ethical standards

This study was deemed exempt by the MIT Committee on the Use of Humans as Experimental Subjects, protocol E-4195.

Keywords

Misinformation; Trustworthiness rating; News Sources; News Headlines.

Appendices

Appendix A: Main analysis and robustness checks

Model 1 in Table 1 reports the main regression model, Model 2 restricts the analysis to participants who did not fail the attention check, Model 3 excludes from the analysis participants who reported they do not share political news on social media, and Model 4 checks whether the labeling effects are moderated by political concordance of the headlines, or political partisanship. The sharing discernment is significantly higher for politically concordant headlines compared to politically discordant headlines for participants in the Fact-checkers condition.

Table 1: Linear regressions with robust standard errors clustered at the participant level.
* p<0.05, ** p<0.01, *** p<0.001

	(1)	(2)	(3)	(4)
	Sharing	Sharing	Sharing	Sharing
True	0.056***	0.060***	0.077***	0.056***
	(0.006)	(0.007)	(0.010)	(0.006)
FC Rated	-0.084***	-0.090***	-0.153***	-0.084***
	(0.017)	(0.019)	(0.023)	(0.017)
LP Rated	-0.042*	-0.029	-0.071**	-0.042*
	(0.17)	(0.019)	(0.024)	(0.017)
FC Unrated	-0.064***	-0.072***	-0.118***	-0.064***
	(0.017)	(0.019)	(0.023)	(0.017)
LP Unrated	-0.035*	-0.025	-0.063**	-0.035*
	(0.017)	(0.019)	(0.023)	(0.017)
True × FC Rated	0.090***	0.091***	0.117***	0.090***
	(0.012)	(0.013)	(0.018)	(0.012)
True × LP Rated	0.052***	0.047***	0.045**	0.053***
	(0.011)	(0.013)	(0.017)	(0.011)
True × FC Unrated	0.045***	0.045***	0.052**	0.046***
	(0.011)	(0.012)	(0.017)	(0.011)
True × LP Unrated	0.033**	0.033**	0.030	0.034**
	(0.010)	(0.011)	(0.016)	(0.010)
Concordant				0.172***
				(0.013)
True × Concordant				-0.030**
				(0.011)
FC Rated × Concordant				-0.032
				(0.018)

LP Rated × Concordant				0.007
				(0.019)
FC Unrated × Concordant				-0.004
				(0.019)
LP Unrated × Concordant				0.008
				(0.019)
True × FC Rated × Concordant				0.057**
				(0.019)
True × LP Rated × Concordant				-0.008
				(0.019)
True × FC Unrated × Concordant				0.001
				(0.019)
True × LP Unrated × Concordant				0.005
				(0.018)
Party				-0.004
				(0.013)
True × Party				-0.018**
				(0.006)
FC Rated × Party				0.019
				(0.018)
LP Rated × Party				0.008
				(0.018)
FC Unrated × Party				0.019
				(0.018)
LP Unrated × Party				-0.000
				(0.018)
True × FC Rated × Party				-0.009
				(0.012)
True × LP Rated × Party				-0.008
				(0.011)
True × FC Unrated × Party				-0.010
				(0.010)
True × LP Unrated × Party				-0.009
				(0.011)
Constant	0.364***	0.360***	0.485***	0.364***
	(0.013)	(0.014)	(0.016)	(0.013)
N	39048	32352	19848	39048

As a robustness check, we have run the main regression with all the subjects (those who failed the attention check, Berinsky, Margolis, and Sances (2014)) and, overall, we found that the results follow the same direction as the main analysis, although the LP Rated and LP Unrated coefficients lose significance, while the FC Rated and the FC Unrated remain significant (see Table 2).

Table 2: Linear regressions with robust standard errors clustered at the participant level. * p<0.05, ** p<0.01, *** p<0.001

	Sharing
True	0.057*** (0.005)
FC Rated	-0.049*** (0.015)
LP Rated	-0.022 (0.015)
FC Unrated	-0.029* (0.014)
LP Unrated	-0.012 (0.015)
True x FC Rated	0.062*** (0.010)
True x LP Rated	0.040*** (0.010)
True x FC Unrated	0.024** (0.009)
True x LP Unrated	0.022* (0.09)
Constant	0.365*** (0.010)
N	53976

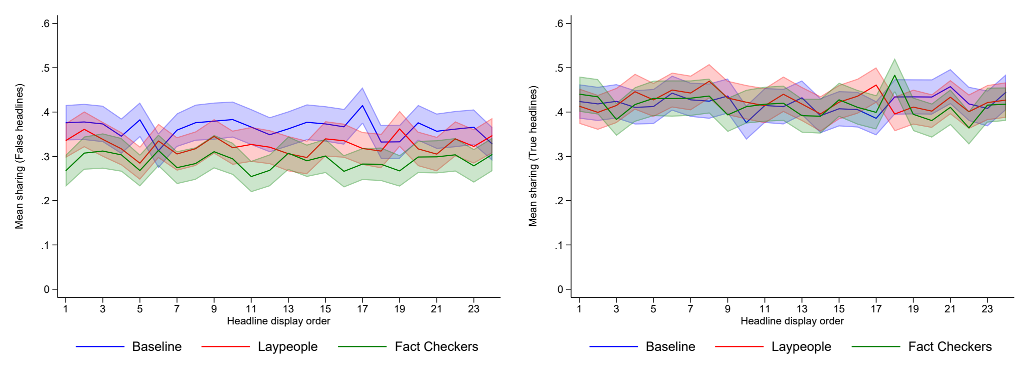
Appendix B: Duration of the effect

In Table 3, we check whether the effect of the intervention decays over time. Specifically, we use the headlines display order as a proxy of time. Model 1 is the main regression model, Model 2 restricts the analysis to participants who did not fail the attention check, and Model 3 excludes from the analysis participants who do not share political news on social media. Overall, we found little evidence that the effect decays over time. Only the interaction Order x True x LP Unrated is significant in two out of three models, suggesting that the indirect effect of laypeople ratings on sharing discernment may slightly decrease over time. Figure 4 plots the mean sharing intentions over time, by treatment and headline veracity.

Table 3: Linear regressions with robust standard errors clustered at the participant level. Significance levels: *** p<0.001, ** p<0.01, * p<0.05

	(1)	(2)	(3)
	Sharing	Sharing	Sharing
True	0.045**	0.040**	0.060**
	(0.014)	(0.015)	(0.020)
FC Rated	-0.085***	-0.106***	-0.147***
	(0.022)	(0.024)	(0.030)
LP Rated	-0.035	-0.029	-0.056
	(0.022)	(0.025)	(0.030)
FC Unrated	-0.073***	-0.088***	-0.139***
	(0.022)	(0.024)	(0.031)
LP Unrated	-0.056*	-0.059*	-0.082**
	(0.022)	(0.025)	(0.031)
True × FC Rated	0.099***	0.113***	0.126***
	(0.024)	(0.027)	(0.034)
True × LP Rated	0.044	0.047	0.047
	(0.024)	(0.026)	(0.034)
True × FC Unrated	0.080***	0.090***	0.107**
	(0.024)	(0.026)	(0.033)
True × LP Unrated	0.078***	0.098***	0.075*
	(0.023)	(0.026)	(0.033)
Order	-0.001	-0.001	-0.000
	(0.001)	(0.001)	(0.001)
Order ×True	0.001	0.002	0.001
	(0.001)	(0.001)	(0.001)
Order × FC Rated	0.000	0.001	-0.000
	(0.001)	(0.001)	(0.002)
Order × LP Rated	-0.001	0.000	-0.001

	(0.001)	(0.001)	(0.002)
Order × FC Unrated	0.001	0.001	0.002
	(0.001)	(0.001)	(0.002)
Order × LP Unrated	0.002	0.003*	0.002
	(0.001)	(0.001)	(0.002)
Order × True × FC Rated	-0.001	-0.002	-0.001
	(0.002)	(0.002)	(0.002)
Order × True × LP Rated	0.001	0.000	-0.000
	(0.002)	(0.002)	(0.002)
Order × True × FC Unrated	-0.003	-0.004*	-0.004
	(0.002)	(0.002)	(0.002)
Order × True × LP Unrated	-0.004*	-0.005**	-0.004
	(0.002)	(0.002)	(0.002)
Constant	0.370***	0.374***	0.491***
	(0.015)	(0.016)	(0.020)
N	39048	32352	19848



(a)

(b)

Figure 4: Mean sharing intention of false news headlines (in panel a) and true news headlines (in the panel b) by headline display order.