Social Truth Queries: Development of a New User-Driven Intervention for Countering Online Misinformation

Madeline Jalbert,[1]*, Morgan Wack[1], Pragya Arya[2], & Luke Williams[3]

[1]University of Washington, Center for an Informed Public, Seattle, WA, United States.

[2]University of Southern California, Department of Psychology, Los Angeles, CA, United States.

[3]Pomona College, Politics Department, Claremont, CA United States.

*Corresponding author. Email: mjalbert@uw.edu

Version: September 2023

***Journal of Applied Research in Memory and Cognition*, 2023**

This is the accepted manuscript prior to copy-editing by the publisher

Abstract

The spread of false information on social media is a growing problem necessitating the development of interventions to reduce its impact. We tested the potential effectiveness of social "truth queries" — user replies that draw attention to truth — as a novel intervention for reducing the impact of false information shared on social media. Participants were shown Tweets containing false information that appeared with user replies containing truth queries (Experiments 1-3), no replies (Experiments 1-3), or user replies unrelated to truth (Experiments 2-3) and asked to judge either the truth of the information contained in the Tweet or their likelihood of sharing it. We consistently found that social truth queries reduce belief in and reported intent to share Tweets containing false information compared to no replies or replies unrelated to truth. The findings suggest the usefulness of truth queries as a simple, flexible, user-driven approach to addressing online misinformation.

Keywords: misinformation; social correction; accuracy nudge; judgment and decision-making; fake news

General Audience Summary

Many individuals wonder what they can do when they see other users post false information online. Fact-checking can take substantive time and effort, and the information may not always be available to perform these fact-checks. People may also be hesitant to directly correct and confront their peers online. We test a novel, alternative intervention for addressing online misinformation: responding to posts containing false information with "truth queries": questions that draw attention to truth or criteria used to judge truth, such as the presence of evidence or the credibility of the information's source. Examples of truth queries include "How do you know this is true?", "Where did you learn this?", and "Do you have an example?". These questions may alert other users to pay more attention to the accuracy of the false information and communicate that the information isn't universally accepted. In a series of three studies, we showed participants Tweets containing false information that appeared with no reply, a reply containing a truth query, or a reply unrelated to truth. We found that the presence of a user reply containing a social truth query consistently reduced participants' belief in and intent to share the posts containing false information compared to when the same posts appeared with no replies or a reply unrelated to truth. We also found that a variety of different types of truth queries were effective, making this approach flexible and adaptable. This research provides initial evidence for the use of social truth queries as an easy-to-implement, non-confrontational, user-driven strategy for addressing misinformation online.

Social Truth Queries: Development of a New User-Driven Intervention for Countering Online Misinformation

The spread of false information on social media is a growing problem necessitating the development of novel interventions to reduce its impact. Corrective messaging is a commonly proposed strategy and can often be effective in reducing belief in false information (Chan et al., 2017; Swire-Thompson et al., 2020; Walter & Murphy, 2018). However, implementing corrective messaging on social media presents unique challenges. Individuals may not trust the legitimacy of the fact-checking services that perform corrections, and social media companies may not be incentivized or have the resources to fact-check themselves (Brandtzaeg & Følstad, 2017). And while corrections made by other users can be effective (Badrinathan & Chauchard, 2023; Bode & Vraga, 2018; Vraga & Bode, 2021), individuals may be hesitant to directly call out their peers.

The short-comings of traditional fact-checking methods have led to the development of alternative strategies that, instead of presenting users with explicit fact-checks, encourage users to consider the truth of information before deciding to share it. As discussed below, these strategies have demonstrated promise for improving the quality of information shared online. Building off these approaches, we test the potential effectiveness of a novel strategy developed through insights from activists in the Global South that utilizes social "truth queries" as a flexible, low-cost, user-driven strategy. Evidence from these tests highlights the potential of truth queries as an innovative and effective tool for countering the spread of online misinformation.

Attention to Truth and Accuracy Nudges

While consuming information online, users may not stop to consider whether it is true. Indeed, evidence suggests that the default assumption is that information is true unless there are

cues present to indicate that a communication may be uncooperative (Grice, 1975; Schwarz, 1994, 1996; Sperber & Wilson, 1986). Additionally, even when truth can be determined, people may instead focus on goals unrelated to accuracy when deciding what to share, such as entertainment and forming social bonds (Hirst & Manier, 2008; Hyman JR., 1994; Hyman Jr., 1999; Marsh & Tversky, 2004). The affordances of social media platforms lead users to seek out positive rewards in the form of likes and comments, which can take priority over the sharing of accurate information. Sharing habits built off these rewards may also make people less sensitive to the consequences of their sharing over time (Ceylan et al., 2023). In other words, there is a disconnect between people's sharing behavior and their judgments of information accuracy (Epstein et al., 2021; Pennycook et al., 2021).

An alternative approach to corrective messaging focuses on interventions that disrupt typical information processing tendencies and cue people to consider truth. Research suggests that giving people a reason to be skeptical and attend to truth can effectively increase critical analysis and reduce subsequent belief in false information (Lewandowsky et al., 2012; Schwarz & Jalbert, 2020). For example, Jalbert, Newman, and Schwarz (2020) found that warning people that some information they were about to encounter was false reduced the impact of viewing that information on later belief. Similarly, drawing attention to truth can also reduce the intent to share false information on social media. In one set of studies, participants who were asked to explain how they knew headlines were true before deciding to share reported lower intentions of sharing fake (but not true) news stories (Fazio, 2020). In another study, respondents asked to judge accuracy before sharing a post reduced their intent to share both true and false headlines, with the impact being greater for the false headlines. Sharing intent decreased even further for

both true and false headlines when participants were additionally asked to provide their rationale for why they believed the headlines were or were not accurate (Jahanbakhsh et al., 2021).

Along the same lines, recent research has also investigated the use of accuracy "nudges" (also called accuracy primes or prompts) to improve the quality of information shared online. These nudges involve asking people to judge the truth of a single headline with the goal of getting users to shift attention to the truth of subsequent information they encounter. They have been found to reduce the intent to share false headlines and increase the quality of information shared (Pennycook & Rand, 2022). Similar to the findings regarding the effect of warnings of false information on later truth judgments (Jalbert et al., 2020), accuracy nudges seem to be effective not necessarily because they increase how much people think about information, but rather because they change what people are thinking about, with increased consideration of the information's veracity (Lin et al., 2022).

Strategies that orient users to consider information truth have advantages over traditional fact-checking approaches because they can be implemented quickly, are scalable, and do not require analysis of the quality of news. Unfortunately, existing strategies also have their limitations: they disrupt the user's normal experience on social media, often involve additional effort on the part of the user, and require that users maintain a focus on accuracy for all posts they encounter, which may be unrealistic in the real world. In addition, strategies integrated into social media platforms require buy-in from social media companies to implement.

Present Strategy: Social Truth Queries

In light of the limitations of existing strategies, organizations have started to develop their own community-driven strategies to combat the spread of harmful information online. One such non-profit organization is the Centre for Analytics and Behavioural Change, which created

Democracy Yethu Kaofela, a project targeting election-related misinformation in South Africa. This project mobilizes community volunteers to respond to misinformation online using techniques developed by a team of in-house dialogue facilitators with the goal of reducing the impact of this misinformation on future viewers. These replies do not require explicit fact-checking and often take the form of questions. For example, a user might reply "how do you know this is true?" or "where did you learn this?".

Inspired by the approach used by Democracy Yethu Kaofela and the success they have observed, we aimed to test key principles behind this intervention. Specifically, we investigated whether the presence of user replies containing questions asking about truth or truth criteria targeted to posts containing false information could reduce other users' belief in and intent to share these posts.

Research on judgment suggests that people utilize a limited set of truth criteria when deciding whether or not a claim is true (Schwarz, 2015, 2018; Schwarz et al., 2016; Schwarz & Jalbert, 2020): Compatibility (Is the claim compatible with other things they believe?), coherence (Is it internally coherent?), credibility (Does it come from a credible source?), consensus (Do others believe it?), and evidence (Is there supporting evidence?). When evaluating whether something is true, people will typically consider some — but not all — of these criteria. In addition to directly drawing attention to truth by explicitly asking if the information is true, we anticipate that people can be cued to consider truth by drawing attention to criteria bearing on truth.

We coin the term "truth queries" to refer to questions that draw attention to these truth criteria or to truth more generally. This represents the first time these types of questions have been defined and systematically tested as a strategy for addressing false information. In our

studies, we created and tested a set of these truth queries that appeared as user replies to social media posts containing false information. For example, a reply drawing attention to the information source might ask, “where did you learn this information?”, while a reply drawing attention to the amount of evidence for the information might ask, “what is the evidence for this?”. Allowing users to utilize an array of truth queries that appeal to different truth criteria creates a flexible approach that does not require a user to post a specific response, but rather choose a response they prefer that matches the context of the correction. Current interventions focusing on attention to accuracy usually ask the same question over and over, which may have a reduced impact over time as people adapt to the question. Here, users may adapt their responses from an array of different truth queries to fit the specific context. In addition to the flexible nature of the response options, this truth query strategy has several advantages over existing attention to accuracy: it can be implemented by social media users themselves, is integrated into the viewer’s normal social media experience, and can be targeted at specific posts containing problematic false information rather than being applied broadly to all types of information.

Social Truth Queries: Beyond Attention to Accuracy

Social truth queries may be effective for reasons beyond merely getting users to consider accuracy. This strategy involves other users sharing replies that express uncertainty or doubt about the content of social media posts, communicating that there may be a lack of social consensus around that information. This information is important given how people look to the beliefs of others as a way to assess accuracy. The perception that a belief is widespread fosters its acceptance — if many people believe something, there must be something to it (Festinger, 1954). Consistent with this idea, previous research has found that people use the comments of other users as cues to the credibility of the information, with comments expressing concerns regarding

information credibility reducing users perceived credibility (Kluck et al., 2019). In addition, people are more likely to share — and less likely to fact-check — information when it has indicators of high social engagement on social media such as likes and shares (Avram et al., 2020). In the absence of these user replies or other cues to truth, people often nod along with the information presented. Truth queries may disrupt these assumptions and communicate to readers that a claim is controversial.

An additional consideration is that, because user replies are often shared directly on the post containing the false information (such as on Twitter), sharing the post also means sharing the attached reply expressing uncertainty. Given the motivation of people to affiliate and receive the approval of others (Cialdini & Goldstein, 2004), people may be less willing to share a post on social media when a comment questioning its truth comes attached to it, separate from whether or not they believe the information in the post to be true.

Present Investigations

We conducted three online studies with Amazon MTurk workers (Experiment 1-2) and Prolific workers (Experiment 3) to investigate the effectiveness of social truth queries — user replies containing questions asking about truth or truth criteria — as a novel approach to reduce belief in and the intent to share false information posted on social media. Our experiments followed the same basic methodology. In all three of our studies, participants (Experiment 1 N = 200, Experiment 2 N = 600, Experiment 3 N = 600) were asked to view a series of Tweets containing false information. These posts appeared either with user replies containing truth queries (Experiments 1-3), user replies unrelated to truth (Experiments 1-2), or no replies (Experiments 1-3). Participants were randomly assigned to judge either the truth of the information contained in each post or how likely they would be to share each post.

Experiment 1

In Experiment 1, we investigated whether posts containing false information were judged to be less true and less likely to be shared when they appeared with user replies containing social truth queries that draw attention to truth and truth criteria, such as evidence or source credibility (e.g., "How do you know that?", "What proof is there that they're doing this?", "Can you share where you learned this?") compared to when they appeared with no replies (for details, see methods).

All stimuli, data, syntax, and supplemental analysis for this experiment and other experiments in this paper can be accessed at https://osf.io/jbcy6/. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in this experiment and other experiments in this paper. In addition, the procedures for all experiments reported in this paper were conducted in compliance with the University of Washington's Institutional Review Board (IRB); when participants signed up for the experiment, they read an information sheet and indicated their agreement to participate.

Methods

Experimental Design

Experiment 1 used a 2 (reply type: truth query or no reply) level within-subjects design, with each participant making either truth or sharing judgments (between subjects). We counterbalanced which half of the Tweets appeared with a truth query reply between participants to control for item effects.

Participants

We used CloudResearch to recruit online workers located in the United States from Mechanical Turk. Participants met CloudResearch's requirements for approved participants and

had completed 100+ HITs with a 95%+ HIT approval rating. We additionally used CloudResearch's filters to select participants that regularly used Twitter. Participants were required to take the survey on a computer (not a tablet or phone). Additionally, as part of the requirements of our funding, participants could not be affiliated with the University of Washington as employees, the family of employees, or as students involved in this particular research. The study was estimated to take 10 minutes or less and participants were paid $1.33 for their time.

Based on a medium effect size of $d = 0.50$, a sample size of 54 would be required to detect an effect of a repeated measures design, with a correlation between repeated measures of .5, $\alpha = .05$, power $(1-\beta) = .95$, and two-tailed, according to G*Power (Faul et al., 2007). We chose to overpower from this and recruit 100 participants per judgment condition (truth or sharing), giving us 200 participants total. For this experiment and all experiments in this paper, we only included the responses from participants who finished the study. In total, 200 participants completed the study ($M_{age} = 38.25$, $SD = 11.80$; 124 male, 72 female, 3 other, 1 prefer not to say), with 100 in the truth judgment condition and 100 in the sharing judgment condition.

Materials

Post Content Creation. We created stimuli in the form of Tweets, which were based on myths that had been fact-checked on Snopes and found to be false. To create the content of the Tweets containing false information, we turned these myths into posts that resembled the types of posts present on Twitter. This content was then normed for truth ($N = 48$) and sharing ($N = 48$) by Mechanical Turk users recruited through CloudResearch. Participants met CloudResearch's requirements for approved participants and had completed 100+ HITs with a 95%+ HIT approval

rating. In this norming, participants rated a series of 20 statements in the form of plain text that they were told were posts made on Twitter. For each, participants judged either truth ("Is the information contained in this above Tweet true or false", unlabeled six-point scale from "Definitely true" (coded as 6) to "Definitely false" (coded as 1)) or sharing ("If you were to see the above Tweet on social media, how likely would you be to share it?", unnumbered 6-point response scale from "Extremely unlikely" (coded as 1) to "Extremely likely" (coded as 6)).

Eight statements from this norming were then selected and made into the format of Tweets. Statements were selected such that they were rated in the middle of truth (median truth ratings 2 or 3 on the 6- point scale). From there, we created two counterbalances out of the eight key Tweets such that half of the claims would appear with truth queries for some participants and the other half would appear with truth queries for other participants. Claims were similar in truth ratings (*M* truth CB 1 = 3.09; CB 2 = 3.01), and sharing ratings (*M* sharing CB 1 = 2.19; CB 2 = 2.31) across counterbalances.

Reply Creation. Next, we created truth query replies to match each of our selected Tweets. These replies were intended to draw attention to the truth of the post or to criteria relevant to truth (such as evidence or information source) without doing any type of fact-checking. Examples include "Can you share where you learned this?" and "What proof is there that they're doing this?".

Tweet Creation. To create realistic-looking Tweets and replies, we selected stock photos and created names and handles from common U.S. first and last names from census data. We checked that the names we selected were not the same as any well-known celebrities and that the handles we created were currently unused. We included a small number of likes (one to four) on

the original post of around half of these Tweets to make the stimuli set look more realistic. Examples of one of these Tweets appearing with their truth query reply can be seen in Figure 1.

Figure 1. Examples of Twitter stimuli used in Experiment 1. These images represent the versions that appeared with the truth query replies. Versions in the no reply condition showed the same main Tweet but were cropped right above the reply.

Procedure.

Twitter Judgments. Participants were told that they would see a series of 12 Tweets (posts made on Twitter) and their responses and were asked to imagine that they came across these Tweets while using Twitter. Participants were randomly assigned to make either truth ratings or sharing ratings (between subjects) for each Tweet.

Participants in the truth rating conditions answered the question "Is the information contained in this Tweet true or false" and made their response on a six-point unnumbered scale with the endpoints "Definitely true" (coded as 6) on the left and "Definitely false" (coded as 1) on the right. Participants in the sharing rating condition answered the question "How likely would you be to share this Tweet online?" on a six-point unnumbered scale with the endpoints "Extremely unlikely" (coded as 1) on the left and "Extremely likely" (coded as 6) on the right.

Participants were additionally informed that some of the Tweets had responses from other users, and that for these Tweets they should consider the Tweet from the original poster when answering the question. They were also told to not search online when completing the studies, and if they were unsure of an answer, to make their best guess.

Participants then saw the 12 Tweets presented: eight key Tweets containing false information (half with a truth query reply and half with no reply) and four fillers containing true information. These Tweets appeared one at a time in random order and participants made their ratings as they appeared. For the eight key Tweets, participants saw one of the two possible counterbalances, such that four Tweets that were presented with truth query replies in one counterbalance appeared with no replies in the other counterbalance, and the four Tweets that appeared with no replies then appeared with truth query replies.

Individual Difference Measures and Demographics. After making these ratings, participants completed a seven-item Cognitive Reflection Test (CRT): The first three items were a reworded version of Frederick (2005) via Shenhav, Rand, and Greene (2012), followed by the four-item CRT by Thomson and Oppenheimer (2016). We included these measures to explore whether individual differences in tendency to utilize intuitive (vs. analytical) processing moderated the impact of truth queries on judgments of truth and sharing.

Finally, participants answered a few demographic questions including age, gender, and political orientation. For political orientation, participants were told "Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal (left) to extremely conservative (right). Where would you place yourself on this scale?". They made their response on a seven-point unnumbered scale with these end-point labels along with "neither liberal nor conservative" as the center label.

Statistical Approach

Data from all studies were analyzed using multilevel modeling with the lme4 package (version 1.1-30) and lmerTest (version 3.1-3) packages in RStudio, which utilize Satterthwaite's degrees of freedom method. We performed separate analyses for the truth and sharing judgments. To investigate the impact of reply condition on truth and sharings ratings, we conducted a mixed effect analysis with reply condition as a fixed factor and item and participant as random factors. We additionally tested for random slopes of reply condition across item and person, but our models failed to converge so we did not include any random slopes in the final models. A parallel analysis using a fixed effect approach is additionally reported for all studies in the Supplementary Materials section 1.

In addition to the analysis reported below, we conducted exploratory analyses investigating whether the effectiveness of truth queries was moderated by individual differences in tendencies to engage in intuitive (vs. analytical) processing (as measured by our seven-item CRT) or by political orientations by adding each item and its interaction with reply condition into our model. Across our three studies, we failed to find evidence that either variable consistently moderated the effect of truth queries, with only one out of twenty possible comparisons reaching significance. Thus, our observed effects appear robust across these measures and we do not discuss them further in this manuscript. However, a full report of these results can be found in the Supplementary Materials sections 2 and 3.

Results

As predicted, Tweets containing false information were judged to be significantly less true when they appeared with a truth query reply, $M = 2.91$, 95% CI [2.51, 3.32], compared to when they appeared without this reply, $M = 3.11$, 95% CI [2.70, 3.51], $t\ (692) = 2.35$, $p = 0.019$, $b = 0.20$ (95% CI [0.03, 0.36]). We also found a significant effect of reply type on sharing ratings. Participants said they would be less likely to share these Tweets when they appeared with a truth query reply, $M = 2.49$, 95% CI [2.11, 2.86], than when they appeared without this reply, $M = 2.70$, 95% CI [2.32, 3.07], $t\ (692) = 2.28$, $p = 0.023$, $b = 0.21$, 95% CI [0.03, 0.39]). See Figures 2 and 3 for density plots of mean participant truth and sharing ratings by reply condition for this Experiment and Experiments 2 and 3.

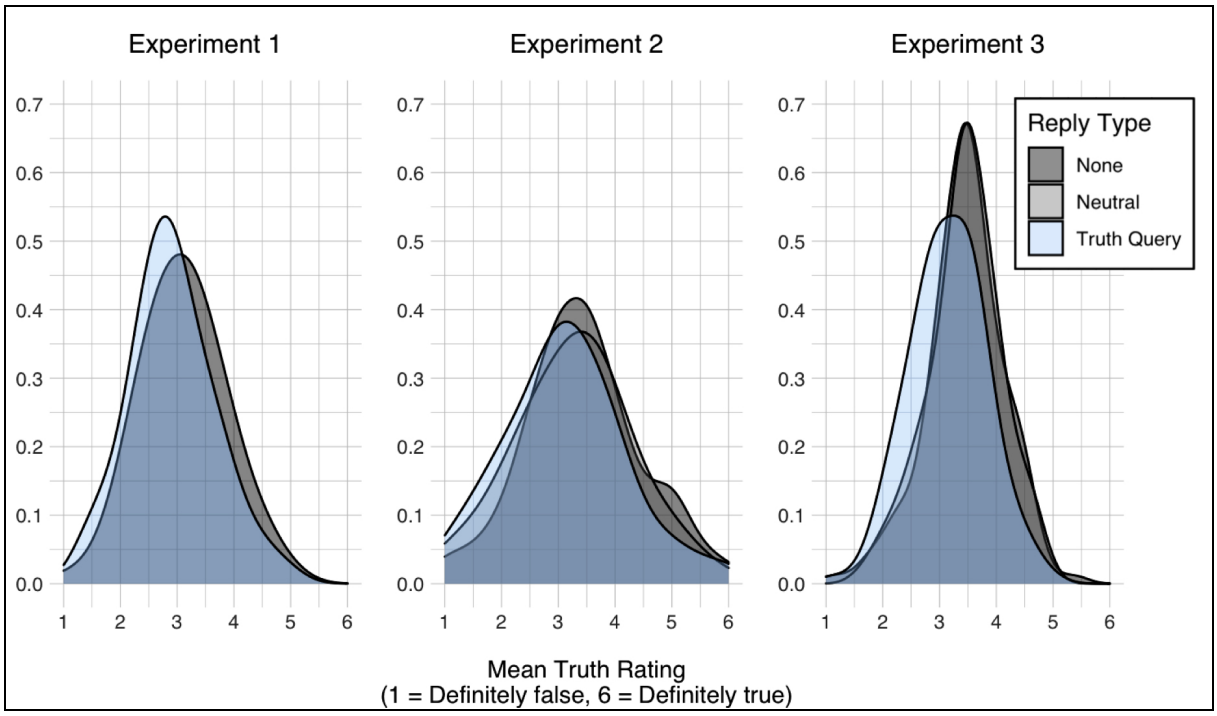

Figure 2. Density plots showing mean truth ratings of Tweets in different reply conditions by participant across Experiments. Ratings were responses to the question "Is the information contained in this above Tweet true or false" made on an unlabeled six-point scale from "Definitely true" (coded as 6) to "Definitely false" (coded as 1).

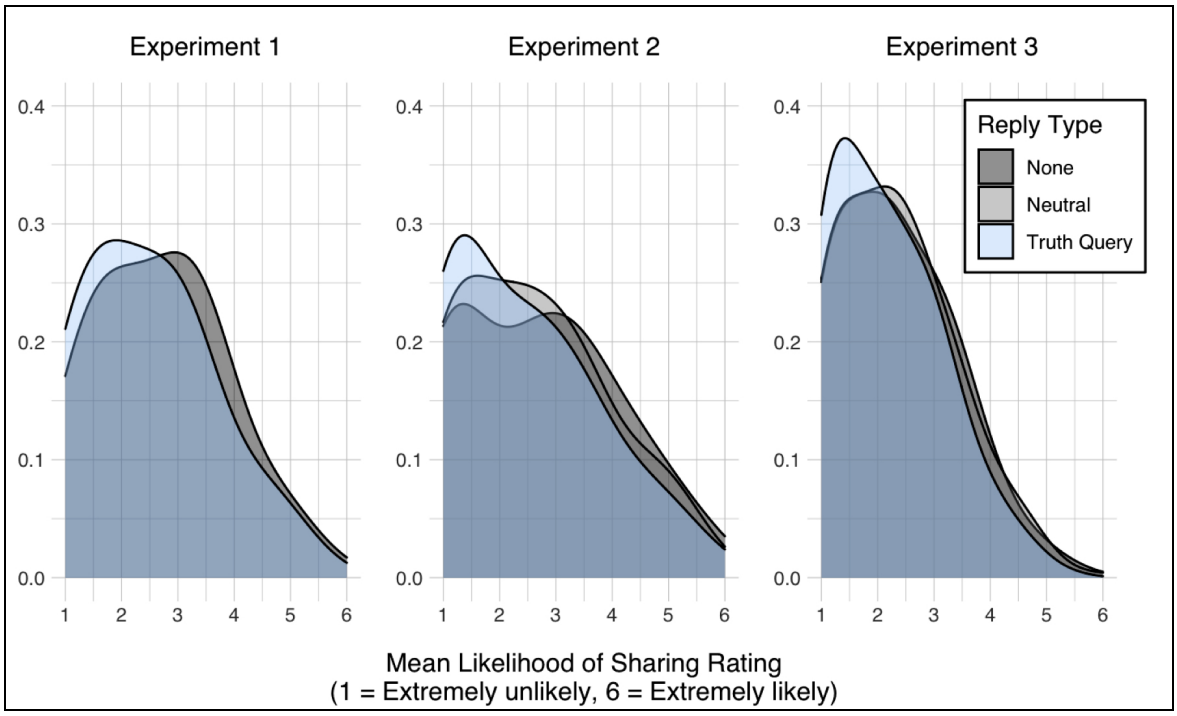

Figure 3. Density plots showing mean sharing ratings of Tweets in different reply conditions by participant across Experiments. Ratings were responses to the question "How likely would you be to share this Tweet online?" on an unlabeled six-point scale from "Extremely unlikely" (coded as 1) to "Extremely likely" (coded as 6).

Experiment 2

One possible concern with interpreting the result of Experiment 1 is whether it was our truth query replies specifically that reduced ratings of truth and sharing, or whether the effect could generally be caused by the mere presence of any reply. Thus, in Experiment 2, we included an additional within-subjects condition in which participants saw the Tweets containing misinformation appearing with replies that were not related to truth or truth criteria — in other words, neutral, non-truth-related replies. For example, in response to a post containing false information related to McDonalds, a neutral reply might be, "I just got home from McDonalds! Didn't get ice cream though". We expected that Tweets that appeared with truth queries would be

judged as less true and less likely to be shared compared to when they appeared with these neutral replies.

Methods

Experimental Design

Experiment 2 used a 3 (reply type: truth query, neutral reply, or no reply) level within-subjects design, with each participant making either truth or sharing judgments (between subjects). We additionally counterbalanced which Tweets appeared with which reply across participants to control for item effects.

Participants

As in Experiment 1, we used CloudResearch to recruit online workers located in the United States from Mechanical Turk that met CloudResearch's requirements for approved participants and had completed 100+ HITs with a 95%+ HIT approval rating. We additionally used CloudResearch's filters to select participants that regularly used Twitter. Participants were again required to take the survey on a computer (not a tablet or phone). The study was estimated to take 10 minutes or less and participants were paid $1.33 for their time.

Based on a small effect size of Cohen's *d* of 0.20, a sample size of 257 would be required to detect an effect of a repeated measures design with three levels, with a correlation between repeated measures of .5, $\alpha = .05$, power $(1-\beta) = .95$, and two-tailed, according to G*Power (Faul et al., 2007). We chose to slightly over-power from this number and recruit 300 participants per judgment condition (truth of sharing), giving us 600 participants total. Overall, 600 participants completed the study ($M_{age} = 38.99$, $SD = 12.47$; 324 male, 268 female, 4 other), 299 in the truth judgment condition and 301 in the sharing judgment condition.

Materials

Our materials were similar to the Tweets used in Experiment 1 with the addition of new neutral (non-truth related) responses to some of the Tweets. Examples of neutral, non-truth-related replies include “I just got home from McDonalds! Didn't get ice cream though” in response to a Tweet containing false information about McDonalds ice cream and “My nephews are in a tag phase right now. It's a good one because they wear themselves out!!” in response to a Tweet containing false information about “tag” being an acronym. Thus, each Tweet had a version with no response, a version with a truth query, and a version with a neutral response. The truth query and neutral responses both came from the same user profile and were identical other than the content of the response itself.

We selected six total Tweets to use from our original norming. As in Experiment 1, we counterbalanced our Tweet-reply pairings such that each Tweet could appear in each reply condition between participants. To do this, we divided our six Tweets into three sets of two Tweets with similar truth and sharing ratings from our earlier norming (truth: *M* CB 1 = 2.86, CB 2 = 3.05, CB 3 = 3.23, sharing: *M* CB1 = 2.54, CB 2 = 2.21, CB 3 = 2.37).

Procedure

The procedure was an exact replication of Experiment 1, except for the specific Tweets that participants rated. While Experiment 1 had 12 Tweets total (four key Tweets with truth queries, four key Tweets with no replies, four fillers), in Experiment 2, participants rated 10 Tweets total, with six key Tweets (two with truth queries, two with neutral replies, and two with no replies) and four fillers. To control for item effects, we randomized which set of two Tweets appeared with truth queries, neutral replies, and no replies between participants.

Statistical Approach

Our statistical approach was identical to Experiment 1.

Results

There was a significant overall effect of reply condition on truth ratings, F (2, 1488) = 8.01, $p < .001$. Posts appearing with truth queries were rated to be less true, $M = 3.12$, 95% CI [2.62, 3.63], than posts with no replies, $M = 3.27$, 95% CI [2.78, 3.78]), t (1488) = -2.00, p = 0.046, b = -0.15 (95% CI [0.00, -0.30]), and posts with neutral replies, $M = 3.43$, 95% CI [2.92, 3.93], t (1488) = 4.00, $p < 0.001$, b = 0.30 (95% CI [0.16, 0.45]). When comparing posts with no replies and posts with neutral replies, those with neutral replies were rated to be significantly more true, t (1488) = 2.00, p = 0.045, b = 0.15 (95% CI [0.00, 0.30]).

In addition, we again found a significant overall effect of reply type on sharing ratings, F (2, 1498) = 5.74, p = .003, with participants responding that they would be less likely to share Tweets when they appeared with a truth query reply, $M = 2.46$, 95% CI [2.00, 2.93], than when they appeared with no reply, $M = 2.63$, 95% CI [2.17, 3.09], t (1498) = 2.18, p = 0.029, b = 0.17 (95% CI [0.02, 0.32]), or with a neutral reply, $M = 2.72$, 95% CI [2.25, 3.18], t (1498) = 3.34, $p < .001$, b = .26, 95% CI [0.11, 0.40]. Participants did not differ in their ratings of how likely they would be to share posts with neutral replies compared to posts with no reply, t (1498) = 1.16, p = 0.248, b = 0.09 (95% CI [-0.06, 0.24]).

Experiment 3

Both Experiments 1 and 2 found that posts containing false information were judged to be less true and less likely to be shared when they appeared with user replies containing truth queries compared to when they appeared with no replies (Experiments 1-2) or with a neutral, non-truth related reply (Experiment 2). In these two studies, we always paired each Tweet with the same truth query that was created specifically for that Tweet. In Experiment 3, we created generally applicable truth query replies that could appear with different Tweets (see Table 1). We

then varied which Tweet appeared with which of our eight total truth queries between participants. This allowed us to test the generalizability of our truth queries and see if there were certain types of truth queries (e.g., ones that appeal to specific truth criteria) that may be driving our observed effects, or whether a variety of truth queries appealing to different truth criteria may be used to reduce the impact of false information spread on social media. Hypotheses and analyses for this experiment were preregistered at https://aspredicted.org/cc6v3.pdf, and this manuscript is consistent with the preregistration.

Methods

Experimental Design

The basic design for Experiment 3 was the same as for Experiment 2. We used a 3 (reply type: truth query, neutral reply, or no reply) level within-subjects design, with each participant making either truth or sharing judgments (between subjects). We additionally counterbalanced which Tweets appeared in which reply condition between participants, and, when Tweets appeared in the truth query condition, we varied which truth queries they appeared with between participants. This allowed us to control for item effects and test the effectiveness of different truth queries across items.

Participants

Participants were Prolific users located in the U.S. who reported using Twitter and had a minimum approval rating of 95%. Based on the same power analysis as in Experiment 2, we again aimed to recruit 300 participants for each judgment condition (truth or sharing), giving us 600 participants total. Overall, 600 participants completed the study (M_{age} = 36.23, SD = 12.22; 288 male, 292 female, 16 other, 4 prefer not to say), 300 in the truth judgment condition and 300 in the sharing judgment condition. Participants were required to take the survey on a computer

(not a tablet or phone). The study was estimated to take 10 minutes or less and participants were paid $1.50 for their time.

Materials

Post Content Creation. Materials were again Tweets that we created to contain false information. This time, in addition to myths fact-checked on Snopes, we also adapted additional false claims that were fact-checked on other online sources or used as materials by Ecker, Lewandowsky, and Chadwick (2020) into the content of our Tweets.

The contents of our Twitter posts were normed on Prolific using participants recruited with the same filters as our main study. We normed 56 claims total, with each participant rating either the first half or last half of the Tweets on truth (Is this information contained in this Tweet true or false, Definitely true; Definitely false), sharing (How likely would you be to share this Tweet online? (Extremely unlikely, Extremely likely) and familiarity (How familiar are you with the information contained in this Tweet?; Not at all familiar; Extremely familiar). We recruited 240 participants total, with each claim being rated on truth, sharing, and familiarity by 39-41 participants.

From this norming, we selected 24 claims to be used as the content of Tweets in our study. Claims were selected to be ambiguous in terms of truth (M = 3.47, ranging between 3.00 and 4.83), be above floor levels on sharing ratings (M = 2.32, ranging between 1.90 and 2.72), and not be too familiar (M = 2.07, range: 1.15 - 3.45). These 24 claims were then divided into three sets of eight claims, with each set having similar truth (M CB 1 = 3.44; CB 2 = 3.48, CB 3 = 3.50), sharing (M CB 1 = 2.31; CB 2 = 2.35, CB 3 = 2.30), and familiarity (M CB 1 = 2.14; CB 2 = 2.02, CB 3 = 2.05) ratings. In Experiments 1 and 2, some of the original posts were presented

with a small number of likes. In this experiment, we kept things simple and did not include likes on any posts.

Reply Creation. For this study, we created a set of eight general truth queries that could be used interchangeably in response to different posts. These claims were created to draw attention to different truth criteria (Schwarz et al., 2016; Schwarz & Jalbert, 2020) or truth generally. A list of these eight truth queries by relevant truth criteria can be found in Table 1.

We had previously divided our 24 Tweets into three sets of eight Tweets based on our norming data. This was so each participant could see one set of eight Tweets with truth query replies, one set of eight Tweets with neutral replies, and one set of eight Tweets with no replies. Which set appeared in each reply condition was randomized between participants.

For the set of eight Tweets that appeared to a participant with truth query replies, each of the eight different truth queries appeared as a reply on one of the eight Tweets. Thus, each participant saw each different truth query on a Tweet exactly once. We also created two different counterbalances of Tweet-truth query pairings for each set to increase the variation in our pairings. This meant that each truth query could appear with two different Tweets from each set, and because there were three sets total, across all participants, each truth query appeared with a total of six different Tweets between-subjects.

Truth query replies were paired with Tweets in these two counterbalances in the following way: To create the first counterbalance, we fully randomized the assignment of the truth queries to the Tweets. For the second counterbalance, we randomized this assignment again, with the condition that each Tweet would have to appear with a truth query from a different category than the first order (e.g., a Tweet couldn't appear with questions related to, say, credibility in both counterbalances). In one instance where the truth query did not follow well

given the wording of the Tweet (Tweet: “Why is Starbucks replacing single-use plastic straws with paper straws in single-use plastic packaging? It makes no sense” Truth query: “Does that make sense given what else you know?”) we used the next randomized option. We again additionally created neutral, non-truth-related replies that match the content of each Tweet. Each Tweet always appeared with the same neutral reply when in the neutral condition.

Procedure

The procedure was again identical to Experiments 1 and 2, except this time participants rated 28 Tweets total: 24 key Tweets (eight with truth query replies, eight with neutral replies, and eight with no replies) and four fillers. Which set of eight Tweets appeared with each reply type was randomized between participants, and participants were additionally randomly assigned to one of the two counterbalances of Tweets-truth query pairings for the eight Tweets they saw with truth query replies.

Table 1. Truth Queries and Their Corresponding Truth Criteria for Experiment 3.

Truth Query	Truth Criteria
1. Does that make sense given everything else you know?	Compatibility: Is it compatible with other things I know?
2. Where did you learn this? 3. How do you know that?	Credibility: Does it come from a credible source?
4. Do other people believe that?	Consensus: Do other people believe it?
5. What evidence is there for that? 6. Is there proof of that?	Evidence: Is there supporting evidence?
7. Why would that be the case? 8. How do you know this is true?	General appeal to truth

Statistical Approach

Our statistical approach for truth and sharing ratings was the same as in Experiments 1 and 2. We also performed additional analysis for Experiment 3 to look at the individual effects of different truth queries. With our counterbalancing, each of our truth queries appeared with six of the 24 total key Tweets used in this study. Thus, for each different truth query, we conducted the same mixed-effects analysis as we did for overall truth and sharing ratings, but limited our comparisons to the ratings of the six Tweets appearing with a specific truth query to those same six Tweets when they appeared with no reply or a neutral reply.

Results

Replicating our prior experiments, we found a significant effect of reply condition on truth judgments, F (2, 6876.7) = 38.59, $p < .001$. Once again, Tweets containing false information were judged to be less true when they appeared with a truth query reply, $M = 3.15$, 95% CI [2.94, 3.36], compared to when they appeared without this reply, $M = 3.45$, 95% CI [3.24, 3.66], t (6876) = -7.61, $p < .001$, $b = -0.30$ (95% CI [-0.38, -0.22]) or when they appeared with a neutral reply, $M = 3.45$, 95% CI [3.24, 3.66], t (6876) = -7.59, $p < .001$, $b = -0.30$ (95% CI [-0.38, -0.22]). Unlike in Experiment 2, we did not find a significant difference in truth ratings for posts appearing with a neutral reply vs. with no reply, t (6878) = -0.016, $p = .988$, $b < 0.01$ (95% CI [-0.08, 0.08]).

Turning to sharing ratings, we again found a significant overall effect of reply type on sharing ratings, F (2, 6875.3) = 15.56, $p < .001$. Participants reported they would be less likely to share these Tweets when they appeared with a truth query reply, $M = 2.17$, CI [2.02, 2.33]), than when they appeared without this reply, $M = 2.33$, 95% CI [2.18, 2.49], t (6875.2) = -4.59, $p < .001$, $b = -0.16$ (95% CI [-0.23, -0.09]) or when they appeared with a neutral, non-truth related

reply, $M = 2.35$, 95% CI [2.19, 2.51], t (6875.3) = 5.04, $p < .001$, $b = 0.18$ (95% CI [0.11, 0.24]). Like in Experiment 2, participants did not differ in their ratings of how likely they would be to share posts that appeared with a neutral reply compared to when they appeared with no reply, t (6875.4) = 0.46, $p = 0.647$, $b = 0.02$ (95% CI [-0.05, 0.08]).

Analysis by Reply

We then looked at the individual impact of each of the eight truth queries used in Experiment 3. Beta statistics and 95% CIs by truth query (compared to both no reply and neutral replies) can be found in Figures 4 and Figure 5 respectively, with full statistics reported in the Supplementary Materials (see Table S1 and S2). These analyses failed to reveal that any specific truth queries were consistently driving the observed effects on truth and sharing judgments. Rather, a broad range of truth queries appeared to be effective in reducing truth and sharing ratings across different Tweets. As the impact of each truth query reply utilizes a different subset of six Tweets, we do not wish to put too much on any one effect size or direct comparison between any of our truth criteria here. However, we do find it encouraging that a broad range of truth criteria demonstrate effectiveness on a variety of Tweets.

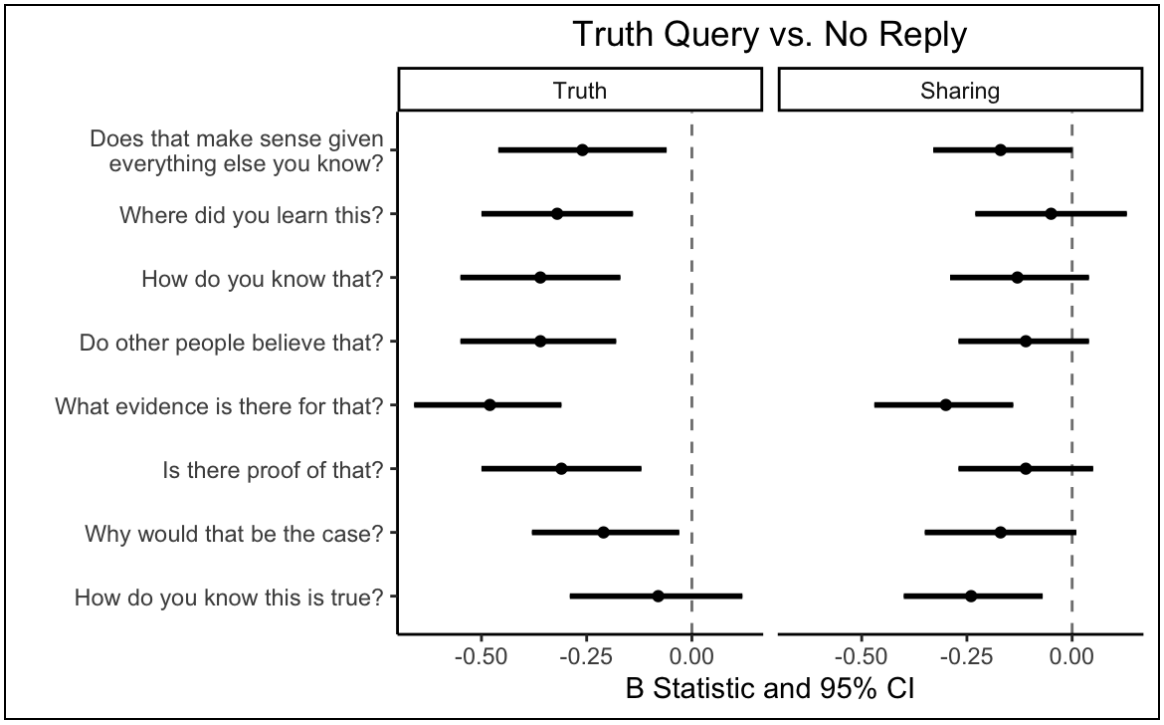

Figure 4. Beta statistics and 95% CIs by truth query for the difference in truth and sharing ratings between posts appearing with each truth query compared to when those same posts appeared with no replies.

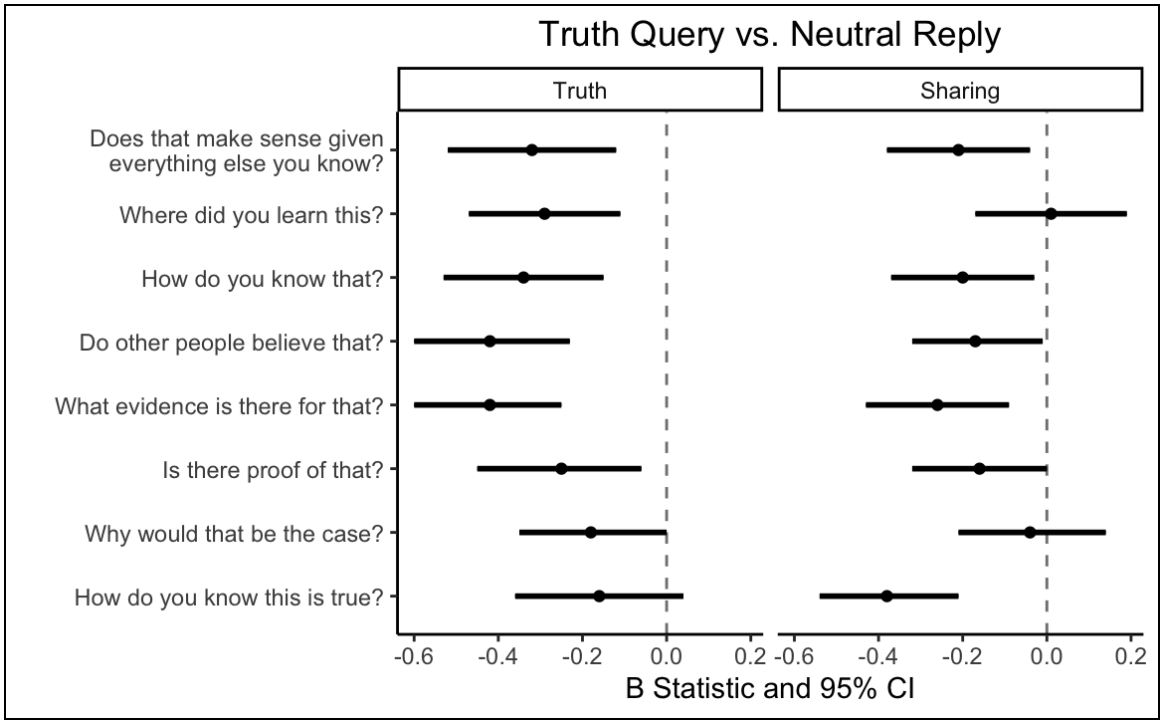

Figure 5. Beta statistics and 95% CIs by truth query for the difference in truth and sharing ratings between posts appearing with each truth query compared to when those same posts appeared with no replies.

Effect Size Analyses Across Experiments

As all of our three studies utilized a similar design, we performed a mini-meta analysis for both truth and sharing ratings across our three studies using Comprehensive Meta-Analysis Software (version 4) to get an overall estimate of effect size. Due to the small number of studies, tau-squared was pooled across studies, following recommendations by Borenstein et al. (2009). A random effects model was used and effect sizes were fixed across sub-groups and corrected for small sample biases (Borenstein et al., 2009). We grouped our effects into two subgroups: the difference in ratings between posts that appeared with truth query replies and posts that appeared with no replies, and the difference in ratings between posts that appeared with truth query replies and posts that appeared with neutral, non-truth-related replies

For truth ratings, the total effect size across all conditions was d = -0.33, 95% CI [-0.47, -0.19]. There was no evidence of a significant difference between the size of the effect when looking at the difference between truth query replies and no replies, d = -0.32, 95% CI [-0.48, -0.11] compared to the difference between truth query replies to neutral replies, d = -0.39, 95% CI [-0.59, -0.16], Q (1) = 0.29, p = .588.

For sharing ratings, the total effect size across all conditions was d = -0.16, 95% CI [-0.21, -0.12]. There was again no evidence of a significant difference between the size of the effect when looking at the difference between truth query replies and no replies, d = -0.17, 95% CI [-0.23, -0.11] compared to the difference between truth query to neutral replies, d = -0.16, 95% CI [-0.22, -0.10], Q (1) = 0.03, p = .873. A forest plot of all effect sizes can be found in the Supplementary Materials (see Figure S1).

Although we did not compare the impact of truth queries on truth ratings vs. sharing ratings directly given we used different measures to assess each, our effect size estimates for truth were consistently about twice as large as those for sharing ratings. One contributing factor to this may be that the claims we chose from norming started at higher ratings of truth at baseline compared to sharing ratings (Experiment 1 M = 3.05 truth vs. 2.25 sharing; Experiment 2 M = 3.05 truth vs. 2.37 sharing; Experiment 3 M = 3.47 truth vs. 2.32 sharing). As participants were already starting closer to floor levels of sharing, there was less room for truth queries to decrease these ratings further.

Discussion

Across three studies, we found a consistent effect of user replies containing truth queries on truth and sharing judgments: Tweets containing false information that appeared with truth queries were judged to be significantly less true and significantly less likely to be shared

compared to Tweets that appeared with no replies (Experiments 1-3) and Tweets that appeared with neutral replies unrelated to truth (Experiments 2-3). These effects were driven by a broad range of truth queries appealing to various truth criteria, such as evidence and social consensus. In Experiment 3, we systematically created and tested a set of eight truth queries appealing to different truth criteria. This allowed us to demonstrate the effectiveness of a variety of truth queries, and see that this effectiveness was not limited to specific Tweet-truth query reply pairings (Experiment 3).

At a more basic level, this work for the first time defines "truth queries" — questions that draw attention to truth or criteria used to access truth (Schwarz, 2015, 2018; Schwarz et al., 2016; Schwarz & Jalbert, 2020) — and demonstrates their potential utility. More broadly, our findings provide compelling evidence for the potential generalizability and flexibility of using social truth queries as a novel strategy for individual users to address a broad range of online misinformation. Importantly, this intervention addresses gaps that exist in current misinformation interventions in several key ways: it is user-driven, flexible, integrated into a user's normal social media experience, and does not require that users directly call out their peers as wrong.

Potential Mechanisms and Applications

Our findings that social truth queries reduced belief in and the intent to share false information are in line with research demonstrating the impact of drawing attention to truth on these outcomes (Fazio, 2020; Jahanbakhsh et al., 2021; Lewandowsky et al., 2012; Pennycook & Rand, 2022; Schwarz & Jalbert, 2020). When people browse and share information online, they may not stop to consider its veracity and assume truth as a default. In addition, other motives, such as entertainment and seeking the approval of peers through the form of engagement may take precedence over accuracy goals. People may also share information — including false

information — out of habit (Bayer et al., 2022; Ceylan et al., 2023). Cues that lead users to shift attention to information truth may shift these patterns of behavior and allow people to catch false information before sharing (Bago et al., 2020) and increase the relevance of information truth as a criterion to consider before sharing.

The effectiveness of the broad range of different truth queries also indicates that direct appeals to truth may not be necessary to be effective. Merely drawing attention to the criteria people use to judge truth — such as the information's source or the presence of evidence — may have a similar effect. An exciting implication of this finding is that existing attention to accuracy interventions, such as accuracy nudges (Pennycook & Rand, 2022) or pausing to consider truth (Fazio, 2020) may be able to utilize a broader range of prompts that appeal to various truth criteria rather than relying on the same question each time. Additional research may further investigate what types of truth queries are maximally effective in different contexts and how much flexibility there is in the form they may take.

We also suspect that truth queries are likely to be effective for reasons beyond attention to accuracy, such as perceptions of social consensus and concerns about self-presentation. As our truth queries were posted by other users, they may also convey that there is a lack of social consensus surrounding the information. Communicating to users that information is disputed among peers may decrease belief in that information as individuals look to what others believe as a way to assess truth (Festinger, 1954). A lack of perceived consensus may additionally reduce the desire to share the post if the truth query reply remains on the original post due to individual self-presentation motives. On the other hand, it is worth noting that our neutral replies did not reduce judgments of truth and intent to share compared to no replies (Experiments 2-3), and in fact, increased perceived truth in Experiment 2. We suspect that perhaps neutral replies —

although unrelated to truth — may act as a form of social proof indicating that others accept the information to be true.

Limitations and Future Directions

Our work was inspired by the work of Democracy Yethu Kaofela, a group run by experienced dialogue facilitators with professional expertise in addressing misinformation. These facilitators train volunteers to respond to targeted areas of misinformation following specific guidelines. Thus, in these initial studies, we chose to test the impact of truth queries only on false information to mirror this approach. However, outside a controlled setting like this, a critical consideration is the potential impact truth queries may have on true information if users — intentionally or unintentionally — respond to posts containing true information instead of false information. Past research utilizing existing attention-to-accuracy interventions has reported mixed findings on their impact on true information. For example, while some research has found that pausing to consider truth only reduced the reported likelihood of sharing false — but not true — news headlines (Fazio, 2020), other research has found that being asked to judge accuracy can reduce the sharing of both true and false news headlines (Jahanbakhsh et al., 2021). Because of the potential harms resulting from reducing belief in credible information, additional research investigating in this area is needed before recommending the implementation of truth queries in settings where they may be used on true information.

In addition, the complex nature of the online information environment lends itself to further considerations. The impact of social truth queries likely depends on contextual factors such as the content of the original post, the characteristics of the users involved, and the presence of additional post engagement. Our studies focused on Tweets containing false information about non-political topics involving unfamiliar users, with minimal engagement outside of our neutral

or truth query replies. The influence of truth queries may look different if, say, that false information pertained to polarized topics, came from a familiar user, or had already received high levels of engagement.

The effectiveness of social truth queries is also likely to depend on the specific affordances of the online environment. For example, as some algorithms prioritize post engagement in deciding what posts users see, one consideration is whether the addition of truth queries could have the unintended consequence of exposing more users to problematic information. Platforms also vary in how post comments are presented to users, and truth queries may be more effective on platforms where they have a higher degree of visibility to other users. Future research may fruitfully investigate the conditions under which social truth queries are maximally effective in reducing the impact of false information.

Conclusion

While additional research is needed to investigate the use of this strategy in different contexts—including those that more closely resemble real online information environments—our initial work provides promising evidence for the effectiveness of social truth queries, a simple, flexible, user-driven strategy for reducing the impact of misinformation online.

Author contributions:

Conceptualization: MJ, MW

Methodology: MJ, MW, PA, LW

Investigation: MJ, MW, PA, LW

Visualization and Data Analysis: MJ, MW, PA

Writing—original draft: MJ

Writing—review & editing: MJ, MW, PA, LW

Acknowledgments

We would like to give our sincere thanks to Jenna-Lee Strugnell and Stef Snel of Democracy Yethu Kaofelo for bringing their innovative dialogue facilitation methods to our attention. These studies would not have been possible without their determined efforts.

The preparation of this article was supported by the University of Washington's Information School's Strategic Research Fund, the University of Washington's Center for an Informed Public, and the John S. and James L. Knight Foundation through funding to the first and second authors.

Open Science and Transparency: All stimuli, data, syntax, and supplemental analysis for this experiment and other experiments in this paper can be accessed at https://osf.io/jbcy6/.

References

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-033

Badrinathan, S., & Chauchard, S. (2023). "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. *The International Journal of Press/Politics*, 194016122311587. https://doi.org/10.1177/19401612231158770

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bayer, J. B., Anderson, I. A., & Tokunaga, R. S. (2022). Building and breaking social media habits. *Current Opinion in Psychology*, *45*, 101303. https://doi.org/10.1016/j.copsyc.2022.101303

Bode, L., & Vraga, E. K. (2018). See Something, Say Something: Correction of Global Health Misinformation on Social Media. *Health Communication*, *33*(9), 1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386

Brandtzaeg, P. B., & Følstad, A. (2017). Trust and distrust in online fact-checking services. *Communications of the ACM*, *60*(9), 65–71. https://doi.org/10.1145/3122803

Ceylan, G., Anderson, I. A., & Wood, W. (2023). Sharing of misinformation is habitual, not just lazy or biased. *Proceedings of the National Academy of Sciences*, *120*(4), e2216614120. https://doi.org/10.1073/pnas.2216614120

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Cialdini, R. B., & Goldstein, N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Ecker, U. K. H., Lewandowsky, S., & Chadwick, M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, *5*(1), 41. https://doi.org/10.1186/s41235-020-00241-6

Epstein, Z., Berinsky, A. J., Cole, R., Gully, A., Pennycook, G., & Rand, D. G. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-71

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-009

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Grice, H. P. (1975). *Logic and Conversation* (pp. 41–58). Brill. https://doi.org/10.1163/9789004368811_003

Hassan, N., Adair, B., Hamilton, J. T., Li, C., Tremayne, M., Yang, J., & Yu, C. (2015). The Quest to Automate Fact-Checking. *Proceedings of the 2015 Computation + Journalism Symposium*.

Hirst, W., & Manier, D. (2008). Towards a psychology of collective memory. *Memory*, *16*, 183–200. https://doi.org/10.1080/09658210701811912

Hyman JR., I. E. (1994). Conversational remembering: Story recall with a peer versus for an experimenter. *Applied Cognitive Psychology*, *8*(1), 49–66. https://doi.org/10.1002/acp.2350080106

Hyman Jr., I. E. (1999). Creating false autobiographical memories: Why people believe their memory errors. In *Ecological approaches to cognition: Essays in honor of Ulric Neisser* (pp. 229–252). Lawrence Erlbaum Associates Publishers.

Jahanbakhsh, F., Zhang, A. X., Berinsky, A. J., Pennycook, G., Rand, D. G., & Karger, D. R. (2021). Exploring Lightweight Interventions at Posting Time to Reduce the Sharing of Misinformation on Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–42. https://doi.org/10.1145/3449092

Jalbert, M., Newman, E., & Schwarz, N. (2020). Only Half of What I'll Tell You is True: Expecting to Encounter Falsehoods Reduces Illusory Truth. *Journal of Applied Research in Memory and Cognition*, *9*(4), 602–613. https://doi.org/10.1016/j.jarmac.2020.08.010

Kluck, J. P. ., Schaewitz, L., & Krämer, N. C. . (2019). Doubters are more convincing than advocates. The impact of user comments and ratings on credibility perceptions of false news stories on social media. *Studies in Communication and Media*, *8*(4), 446–470.

https://doi.org/10.5771/2192-4007-2019-4-446

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*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lin, H., Pennycook, G., & Rand, D. G. (2022). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, *230*, 105312. https://doi.org/10.1016/j.cognition.2022.105312

Marsh, E. J., & Tversky, B. (2004). Spinning the stories of our lives. *Applied Cognitive Psychology*, *18*(5), 491–503. https://doi.org/10.1002/acp.1001

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. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, *13*(1), 2333. https://doi.org/10.1038/s41467-022-30073-5

Schwarz, N. (1994). Judgment in a Social Context: Biases, Shortcomings, and the Logic of Conversation. In *Advances in Experimental Social Psychology* (Vol. 26, pp. 123–162). Elsevier. https://doi.org/10.1016/S0065-2601(08)60153-7

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation* (pp. xi, 112). Lawrence Erlbaum Associates Publishers.

Schwarz, N. (2015). Metacognition. In M. Mikulincer, P. R. Shaver, E. Borgida, & J. A. Bargh (Eds.), *APA handbook of personality and social psychology, Volume 1: Attitudes and*

social cognition. (pp. 203–229). American Psychological Association. https://doi.org/10.1037/14341-006

Schwarz, N. (2018). Of fluency, beauty, and truth: Inferences from metacognitive experiences. In *Metacognitive diversity: An interdisciplinary approach* (pp. 25–46). Oxford University Press. https://doi.org/10.1093/oso/9780198789710.001.0001

Schwarz, N., & Jalbert, M. (2020). When (Fake) News Feels True. In C. Mc Mahon (Ed.), *Psychological Insights for Understanding COVID-19 and Media and Technology* (1st ed., pp. 9–25). Routledge. https://doi.org/10.4324/9781003121756-2

Schwarz, N., Newman, E., & Leach, W. (2016). Making the truth stick & the myths fade: Lessons from cognitive psychology. *Behavioral Science & Policy*, *2*(1), 85–95. https://doi.org/10.1353/bsp.2016.0009

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*, 423–428. https://doi.org/10.1037/a0025391

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Harvard University Press.

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Vraga, E. K., & Bode, L. (2021). Addressing COVID-19 Misinformation on Social Media

Preemptively and Responsively. *Emerging Infectious Diseases*, *27*(2), 396–403. https://doi.org/10.3201/eid2702.203139

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.146756