

# Emotion- versus Reasoning-based Drivers of Misinformation Sharing:

A field experiment using text message courses in Kenya\*

Susan Athey<sup>1</sup>, Matias Cersosimo<sup>1</sup>, Kristine Koutout<sup>1</sup> and Zelin Li<sup>2</sup>

<sup>1</sup>Graduate School of Business, Stanford University

<sup>2</sup>Sloan School of Management, Massachusetts Institute of Technology

June 2023

---

\*We would like to acknowledge the team at First Draft News, now Information Futures Lab, Claire Wardle, Shaydanay Urbani, and Rory Smith for developing the text message courses in this study. We would also like to acknowledge Tommy Shane and Laura Garcia for their contributions to early discussions of the project. We gratefully acknowledge the Golub Capital Social Impact Lab and the Africa Infodemic Response Alliance at the World Health Organization for their financial support of this study. This experiment was preregistered with the AEA RCT Registry under AEARCTR-0009721 and reviewed by the Stanford Institutional Review Board under protocol 57430.

## Abstract

Two leading hypotheses for why individuals unintentionally share misleading information online are that 1) they are unable to recognize that a post contains misinformation, and 2) they make impulsive, emotional sharing decisions without thinking about whether a post contains misinformation. The strategies to counter each of these drivers of misinformation sharing differ by the techniques that they are designed to address. We categorize techniques according to whether they use misleading reasoning to make recognizing misinformation more difficult (reasoning-based) or manipulate emotions to encourage impulsive sharing decisions (emotions-based). To learn whether interventions designed to counter reasoning- or emotion-based techniques are more effective or whether the approaches are complementary, we evaluate three distinct versions of a low-cost and scalable five-day text message educational course. We assess the impact of the courses in a field experiment with approximately 9,000 participants in Kenya. We measure outcomes using a pre-post survey design that elicits intentions to share and find that all treatment courses work, decreasing misinformation sharing 28% on average relative to no text message course. The treatment designed to counter emotion-based techniques, the “Emotions” course, is more effective than teaching about reasoning-based techniques either alone in the “Reasoning” course or in combination with emotion-based techniques in the “Combo” course. Moreover, the Emotions course performs best on misinformation posts that use emotional manipulation, and does no worse than the Reasoning or Combo courses on misinformation posts that use reasoning-based techniques. In a follow-up experiment approximately two months later, 88% of the treatment effect of the three courses on misinformation sharing persists.

## 1 Introduction

The unintentional sharing of false or misleading information—misinformation—poses a growing threat to public health (Ho et al., 2022), democratic institutions (Berlinski et al., 2021), and other domains where beliefs gleaned from media influence decision-making. The proliferation of misinformation has shifted attention from debunking individual pieces of misinformation to educating social media users more broadly about the techniques used to manipulate users online (van der Linden, 2022). While reasoning-based techniques, such as false dichotomies or out-of-context images, have been the focus of interventions to date,<sup>1</sup> emotion-based techniques, such as the use of emotionally charged language in posts, have been shown to be important to the spread of misinformation (Brady et al., 2017; Pröllochs et al., 2021) and to the likelihood that misinformation is believed (Martel et al., 2020).

We make the distinction between reasoning- and emotions-based techniques because strategies to counter each interfere at different stages in the user’s decision-making process. When a user sees a social media post they want to share, they go through a two-stage decision-making process. In the first stage, the user decides whether to share the post immediately or to proceed to the second stage to evaluate whether sharing the post is a good idea, which includes assessing whether it contains misinformation. Emotional language in the post can

---

<sup>1</sup>See Compton et al. (2021); Lewandowsky and Van Der Linden (2021); van der Linden (2022) for reviews.

trigger an impulsive sharing decision at this stage, so strategies to counter emotion-based techniques aim to tame this impulse. In the second stage, conditional on not sharing in the first stage, the user evaluates whether the post contains misinformation and subsequently makes an informed sharing decision. Strategies to counter reasoning-based techniques aim to provide the user with the tools to evaluate posts in this second stage. While either strategy might be effective alone, combining them may be particularly powerful because they intervene at different stages of the decision-making process. By estimating the impact of teaching reasoning-based techniques, emotion-based techniques, and both together, we can learn about the motivations driving users to share misinformation and refine future interventions accordingly.

To study these questions, we collaborated with experts at the nonprofit organization First Draft News to adapt a low-cost, five-day text message course aimed at countering misinformation sharing. In a field experiment that recruited Facebook users in Kenya, we tested three distinct versions of the text message course that teach users to counter either (i) *reasoning*-based techniques, (ii) *emotion*-based techniques, or (iii) both. Participants are randomly assigned to receive one of the three treatment courses—“Reasoning,” “Emotions,” or “Combo”—or one of two baselines. The “No-course” baseline receives no course between the pre- and post-survey; the “Facts” baseline is exposed to daily facts about misinformation but no technique-based education. We evaluate the courses with two online surveys, one conducted prior to the intervention and one conducted after, where each survey measures intention to share for a set of posts including non-misinformation and misinformation posts. The misinformation posts are categorized as “Reasoning” posts, “Emotions” posts, or “Combo” posts depending on the techniques employed.<sup>2</sup> We follow up with participants seven to eleven weeks after completing the course to measure long-run effects.

We study two outcomes, misinformation sharing and “discernment,” where the latter is a composite that weighs misinformation sharing negatively, and non-misinformation sharing positively but with half the weight. Discernment is a useful outcome because many interventions decrease all sharing, while we are particularly interested in interventions that differentially affect misinformation sharing. We find that all treatment courses effectively and persistently reduce misinformation sharing and increase discernment relative to baselines, with the Emotions course being the most effective. Perhaps surprisingly, teaching reasoning-based techniques in addition to emotion-based techniques provides no added benefit. Furthermore, the Emotions course exhibits cross-technique effectiveness, performing no worse than the Reasoning or Combo courses on Reasoning posts while outperforming the Reasoning and Combo courses on Emotions posts. This result suggests that the strategies taught in the Emotions course generalize to posts that do not contain emotional content.

On average, the treatment courses reduce misinformation sharing by 18 percentage points (p.p.), or 28% of the baseline sharing rate among those not exposed to a text message course. The treatment courses also reduce misinformation sharing by 10 p.p. compared to the Facts baseline. Since the Facts baseline makes the topic of misinformation salient, finding a treatment effect over this baseline supports the hypothesis that the treatment courses

---

<sup>2</sup>We included both specific techniques used as examples in the courses and techniques of the same type, either reasoning- or emotion-based, that were not used as examples in the course to avoid “teaching to the test.”

operate through their content rather than just salience. The effect of the treatment courses persists in the long-run, reducing misinformation sharing 9 p.p. more than the Facts baseline after approximately two months, or 88% of the short-run effect. We find similar effects on discernment, although the treatment effects are smaller, partially because the treatment courses also decrease non-misinformation sharing, though the impact on misinformation sharing is larger.

In addition to randomizing participants into one of the five text message course groups, we independently randomize participants to see an “accuracy nudge,” a misinformation intervention that prompts users to evaluate the accuracy of a post. [Pennycook et al. \(2021\)](#) find that such a prompt changes sharing behavior in a field experiment on Twitter, making it the only misinformation intervention to our knowledge that has been evaluated using on-platform behavior as opposed to survey outcomes. We implement a “high-dosage” version of an accuracy nudge by asking participants to rate the accuracy of *each* post before making the sharing decision for that post in the pre- and post-surveys. Participants *not* assigned to see the accuracy nudge are instead first asked about their intention to share on a set of posts, before rating the accuracy of the set of posts.

The inclusion of the accuracy nudge allows us to benchmark the effectiveness of the text message courses relative to an intervention that has been evaluated using on-platform behavior ([Pennycook et al., 2021](#)). As we discuss in more detail below, the major limitation of our study (that it shares with almost all studies on misinformation sharing) is its reliance on survey outcomes. By measuring the effectiveness of the accuracy nudge alongside our treatment courses on survey outcomes, we can make internally valid comparisons to this externally validated intervention.

The accuracy nudge works as expected in our experiment, reducing misinformation sharing by 7 p.p., but not as well as the treatment courses (18 p.p.). Moreover, when comparing the treatment effect of the accuracy nudge for individuals assigned to one of treatment courses to the effect of the accuracy nudge for participants assigned to baseline interventions, we can rule out large differences in the treatment effect of the accuracy nudge, suggesting that any complementarity is small at best. We interpret these results as providing evidence that the effectiveness of the treatment courses on survey outcomes likely translates to changing on-platform misinformation sharing.

The external validity of our experiment is impacted by the representativeness of our experimental sample relative to the population to which this intervention would be applied if “scaled-up;” the naturalness of the experimental setting; and the correspondence of our measured outcome to the desired outcome in the population of interest.<sup>3</sup> First, our experiment recruits participants using advertisements on social media in Kenya where the prevalence of misinformation has been particularly concerning,<sup>4</sup> making it a prime target for a scaled-up version of the text message courses. Within the 13 million Facebook users in Kenya,<sup>5</sup> we

---

<sup>3</sup>We reference the SANS (Selection, Attrition, Naturalness, Scaling) transparency conditions in [List \(2020\)](#) in our discussion of external validity.

<sup>4</sup>The Africa Infodemic Response Alliance at the World Health Organization has highlighted the prevalence of misinformation on social media in Kenya ([Nguyen and Cecchini, 2021](#))

<sup>5</sup>Referenced from <https://www.statista.com/statistics/1029203/facebook-user-share-in-kenya/> on May 1,

recruit a sample of 9,000 users selected on being active in July 2022 and responding to an ad to do a study for mobile airtime payment. In a scaled-up version that targeted all Facebook users in Kenya, participants may be recruited through Facebook ads and paid, as we do, but other likely implementations include a public information campaign or public service announcements on social media.<sup>6</sup> So, although we observe few systematic patterns in attrition rates with the observables in our sample, selection and attrition patterns may both differ in a scaled-up version.

Our implementation of the text message courses reflects how a scaled-up version could work, making our study more natural and increasing its external validity. As in our experiment, a likely implementation of the text message course would have participants enter their phone number (on a web page as opposed to in a survey) to receive daily text messages via a messaging app (our study gave the choice of MMS or WhatsApp) for five days. Explicit incentives, as we offer in our experiment, would increase take-up of the course, especially among those who are unknowingly misinformation sharers and so may not know that they are likely to benefit from an intervention. A school, government body, or company might also require or incentivize (e.g., with prizes or lotteries) individuals to participate in the text message course.

Arguably the most important limitation to the external validity of our approach is that we measure outcomes through surveys, whereas the outcomes of interest are on-platform behavior. While surveys are the most common measurement method in the literature to date (with the notable exception of [Pennycook et al. \(2021\)](#)), self-reported measures also have known limitations such as experimenter demand effects. We address this shortcoming in three ways. First, as discussed above, we use the accuracy nudge as an internally valid benchmark to an externally validated intervention. Second, our survey measures are separated in time and context from the text message courses. Further, we measure long-run outcomes to evaluate the consistency and duration of results, as experimenter demand effects are likely to be most pronounced directly after interacting with the associated course. Third, we use qualitative questions in the follow-up survey to determine whether participants *report* implementing techniques from the course in their social media behavior and whether treated participants do so more than the baseline. The performance of the courses relative to the accuracy nudge, the persistence of the effect of the treatment courses, and the reported behavior in the follow-up survey support the conclusion that the treatment courses' effectiveness like translates to changes in on-platform behavior.

We make three contributions in this study. First, we evaluate interventions to counter reasoning- versus emotions-based techniques versus both. In addition to the new evidence we provide on the relative effectiveness of strategies to counter emotion-based techniques, the categorization into reasoning- and emotion-based techniques based on strategies that interfere at different stages of the user's decision-making process is novel. We discuss the only other study to our knowledge ([Roozenbeek et al., 2022](#)) that randomly assigns either a reasoning- or an emotion-based technique to participants in the next section. Second, we test a novel intervention type: the first text message course for treating misinformation sharing

---

2023.

<sup>6</sup>For example, [Guess et al. \(2020\)](#) evaluate a campaign that showed a list of digital media literacy tips at the top of Facebook news feeds in the U.S. and India.

to our knowledge.<sup>7,8</sup> Third, even though online misinformation is a global phenomenon, we conduct one of the first evaluations of misinformation interventions outside of the U.S. and high-income countries (with the notable exceptions of [Offer-Westort et al. \(2021\)](#) and [Harjani et al. \(2023\)](#)). Other studies that use survey experiments recruit participants globally ([Arechar et al., 2022](#); [Basol et al., 2021](#); [Gavin et al., 2022](#); [Roozenbeek and Van Der Linden, 2019](#); [Roozenbeek et al., 2022](#)) but do not focus on low-income countries in particular. In the next subsection, we place these contributions in the context of the related literature.

## 1.1 Related Literature

The text message course we test in this study is categorized as an “inoculation” intervention. Inoculation theory can be traced back to as early as the 1960s when [McGuire \(1961\)](#) developed the theory to understand persuasion-resistance with an analogy similar to that of a vaccine. The theory has only recently been utilized in the context of misinformation, where the first set of literature primarily studied inoculation in the context of “prebunking” climate change misinformation ([Cook et al., 2017](#); [Van der Linden et al., 2017](#)). This study contributes to the second wave of inoculation interventions focused on technique-based approaches, as opposed to “issue-based” approaches that treat one piece of misinformation at a time ([Lewandowsky and Van Der Linden, 2021](#)).

The most relevant work to this paper evaluates one of two interventions that teach both reasoning- and emotion-based techniques used to mislead users in online posts: the Bad News game<sup>10</sup> and the Inoculation Science videos.<sup>11</sup> Bad News is an online game in which players role-play as a fake news creator attempting to amass followers while maintaining credibility. The game takes approximately 15 minutes to complete by earning six badges corresponding to different techniques used to spread misinformation. One of the badges is for emotional language, while others focus on reasoning-based techniques like impersonation and conspiracy theories. The video series created by Inoculation Science similarly includes a video on emotional manipulation and five other short videos covering reasoning-based techniques like false dichotomies or ad hominem attacks.

In a series of papers starting with [Roozenbeek and Van Der Linden \(2019\)](#), the Bad News game has been shown to decrease susceptibility to misinformation ([Basol et al., 2020](#); [Roozenbeek et al., 2021](#)), including up to 13 weeks after playing the game with regular reminders ([Maertens et al., 2021](#)), although these results did not replicate in India ([Harjani et al., 2023](#)). The general setup in these papers is that a pre- and post-survey embedded in the game asks participants to rate the reliability, manipulativeness, and/or accuracy of posts. Then, they use a convenience sample of people who play the game online and complete the surveys to evaluate change in survey outcomes before and after playing the game. [Basol et al. \(2020\)](#)

---

<sup>7</sup>[Offer-Westort et al. \(2021\)](#) send text messages through a chatbot with all messages delivered in the same session.

<sup>8</sup>Text message courses have proven to be useful and effective in stimulating positive behavioral changes in previous studies ([Armanasco et al., 2017](#)).<sup>9</sup>

<sup>10</sup><https://www.getbadnews.com/books/english/>

<sup>11</sup><https://inoculation.science/inoculation-videos/>

uses Prolific instead of a convenience sample so they can evaluate the performance of the game relative to a control group that plays Tetris. [Martel et al. \(2020\)](#) evaluate the long-term effectiveness of the Bad News games and found that the inoculation effect remains stable for at least 3 months if the participants were assessed regularly after the intervention. Without regular testing, however, the authors found the long term effects to decay significantly over the period of two months.

In [Roozenbeek et al. \(2022\)](#), the authors evaluate each Inoculation Science video on technique recognition, confidence, trustworthiness, and sharing discernment. Participants on Prolific view one of the six videos, then [Roozenbeek and Van Der Linden](#) show them posts that either do or do not include the technique from the relevant video (i.e., the emotional manipulation video is evaluated using posts that either include emotional manipulation or are non-misinformation). [Roozenbeek et al. \(2022\)](#) find that all videos improve most outcomes, but they do not explicitly discuss the differences they observe between the videos, likely because there are differences between the posts used to evaluate each video. Their results do not show that the emotional manipulation video does best in any of the outcomes they assess (including willingness to share, the metric we use in this paper, where they find no differences between the videos). [Roozenbeek et al.](#) also evaluate two of the videos on YouTube (emotional manipulation and false dichotomies) by randomly showing one of two of the videos to users as ads on YouTube. A subset of those users completed a one-question survey about whether the technique they learned about in the video was present in the headline they were shown. Compared to a control who did not see the videos, users were more likely to correctly identify the technique.

In another study with similarities to our own, [Offer-Westort et al. \(2021\)](#) conduct an adaptive experiment in sub-Saharan Africa (Kenya and Nigeria) to evaluate 11 interventions and determine both the best treatment overall and heterogeneity in optimal treatment assignment. Like our study, [Offer-Westort et al.](#) recruits participants through Facebook ads, but the interventions they test are deployed through a chatbot. [Offer-Westort et al.](#) also use a pre-post survey design, but because the chatbot delivers the intervention instantaneously, they measure intentions to share in the same session in which the intervention is delivered. The “Facebook tips” intervention that [Offer-Westort et al.](#) test, in which the chatbot provides users with strategies to counter misinformation, is similar to a text message course.<sup>12</sup> Our intervention differs from theirs in that our course is delivered over multiple days (compared to in the same session) through either text or WhatsApp (as opposed to Facebook Messenger) and contains some interactive questions.

[Pennycook et al. \(2021\)](#) implement an accuracy nudge by sending an unsolicited direct message to a subset of Twitter users who had recently engaged with untrustworthy news sites that contained one post with a question: “How accurate is this headline?” Users did not need to engage with the question; instead, [Pennycook et al.](#) estimate an intent to treat on exposed users and find increased discernment of misinformation in the 24 hours following exposure. [Offer-Westort et al. \(2021\)](#) include a similar accuracy nudge in their experiment—one post with a question about the post’s accuracy.

---

<sup>12</sup>Facebook tips were also developed by First Draft News, which developed the treatment courses in this study.

Outside of the misinformation context, educating participants about how an organization might be manipulating them has been employed in other contexts. For example, [Bryan et al. \(2016\)](#) use a randomized experiment to evaluate an intervention that teaches adolescents about how the food industry manipulates consumers; it was shown to improve the diets of the participants relative to several baselines. [Bryan et al. \(2016\)](#) interprets the finding as evidence that adolescent motivation around autonomy provides a symbolic reward (“feeling like a... respect-worthy person acting in accordance with important values shared with peers”) that competes with the short-term temptation of unhealthy eating, a theory that has grounding in the neuroscience literature ([Telzer et al., 2014](#)). In our study, the symbolic reward around avoiding manipulation by an organization might compete against the temptation of succumbing to fear, shock, or outrage and sharing a post.

## 2 Experimental Design

The goal of this experiment is to study the effect of different text message courses on misinformation sharing and discernment. [Figure 1](#) diagrams the experimental design. We recruited participants through Facebook ads to complete a five-day text message course, plus a pre- and post-survey, for a mobile airtime payment. Participants who clicked an ad were directed to the pre-survey on Qualtrics, where they were first randomized on whether they would see an accuracy nudge in the pre- and post-survey. Conditional on completing the pre-survey, participants were randomized into one of the text message course interventions and enrolled in the course. Participants in all course interventions, except the No-course baseline, received one text message a day for five days starting on the day they completed the pre-survey.<sup>13</sup> On the last day of the course, participants received a link to complete the post-survey on Qualtrics. Participants who completed the post-survey were paid KSH 500 (about \$4 in U.S. dollars) in mobile airtime. Seven to eleven weeks later, participants who completed the post-survey were randomized on whether they would see a prime in the text message recruiting them to a follow-up survey. Participants who completed the follow-up survey were paid KSH 350 (about \$3 in U.S. dollars).

In this section, we first detail the Stage 1: Main Experiment that encompasses recruitment to the pre-survey through the post-survey. There are two independent randomizations in this experiment: the accuracy nudge in the surveys and the text message course interventions. There are five treatment course interventions; hence, the experiment is a five (course interventions) by two (accuracy nudge) factorial design. For the majority of our analysis, we aggregate over exposure to the accuracy nudge. We exploit the factorial design in [Results section 3.4](#) to learn whether the treatment courses are more or less effective than the accuracy nudge at decreasing misinformation sharing, and whether there are incremental effects of combining the accuracy nudge with the treatment courses.

We then detail the Stage 2: Follow-up Experiment, in which there is a third independent randomization into whether participants are exposed to a prime. The follow-up survey

---

<sup>13</sup>Participants in the No-course baseline received the Combo course after the post-survey to ensure we fulfilled our recruitment promise of a text message course.



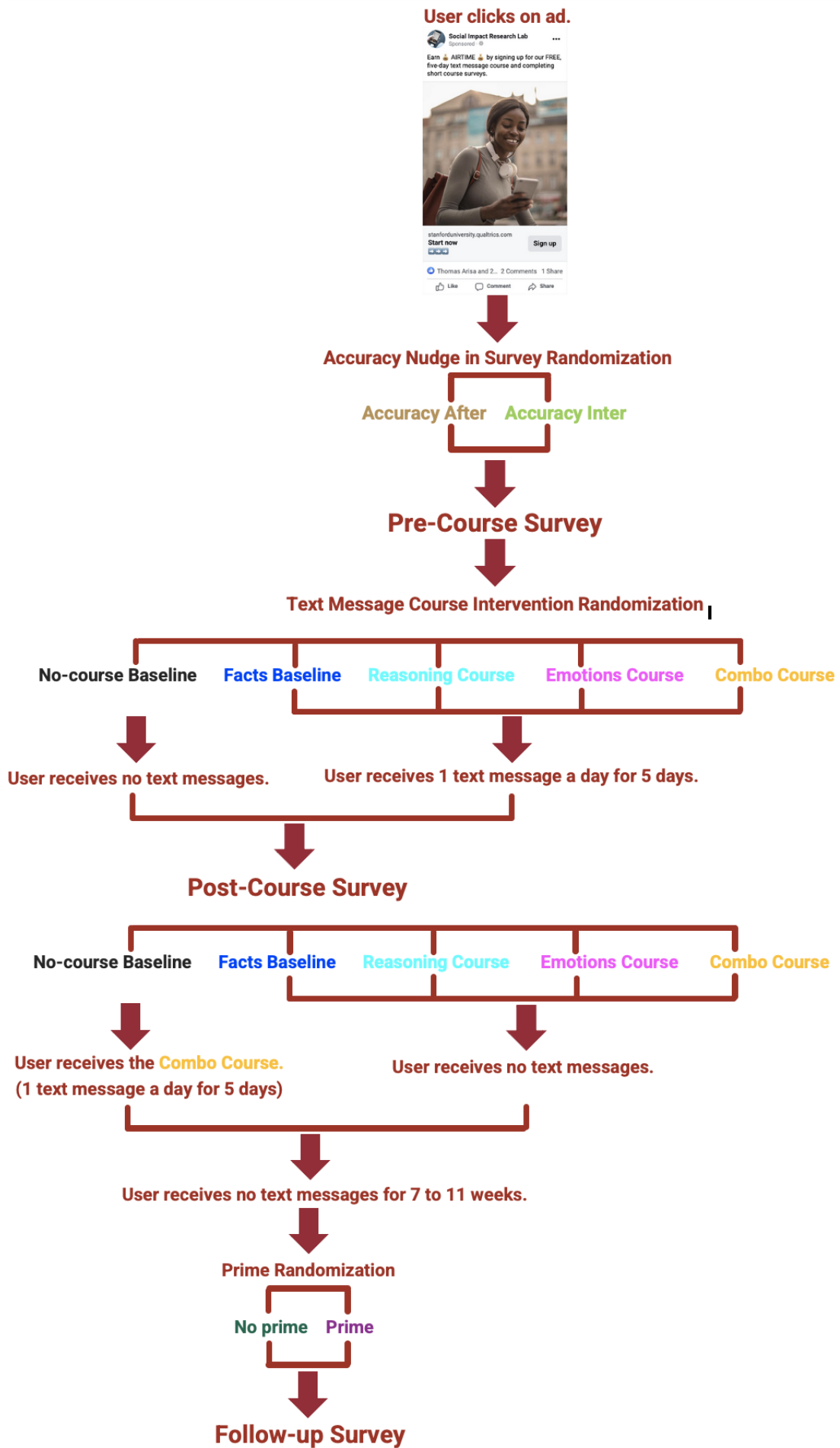


Figure 1: Experiment Design

contains no accuracy questions, so there is no accuracy nudge for any group. Lastly, we discuss implementation of the experiments.

## 2.1 Stage 1: Main Experiment

We recruited our sample from English-speaking Facebook users in Kenya who were at least 18 years old.<sup>14</sup> Ads ran on Facebook and Messenger, where users saw the opportunity to “Earn AIRTIME by signing up for our FREE, five-day text message course and completing short course surveys.” Clicking on an ad led participants to the pre-survey. Participants were told that those who completed the pre-survey, text message course, and the post-survey would receive KSH 500 in airtime, conditional on correctly answering all attention checks (example of an attention check question is included in Online Appendix D).<sup>15</sup> Appendix Table C1 shows our participant funnel from impressions to completion, with associated ad costs (\$1.21 for post-survey completion only, and about \$2 for post-and follow-up survey completion). Online Appendix A contains a sample of the ads we ran.

Participants enrolled in the course at the end of the pre-survey by providing their phone number and choosing whether to receive the text message course via MMS or WhatsApp, a messaging platform with 97% penetration in Kenya.<sup>16</sup> Upon enrollment, participants received the first of five daily text messages delivered at the same time (the time of enrollment) each day, unless they were assigned to the No-course baseline. Each day’s message was equal to or less than 1200 characters long. We next describe the five text message course interventions.

### 2.1.1 Text Message Course Interventions

First Draft News, a nonprofit organization focused on protecting communities from misinformation, has been a leader in developing interventions to combat misinformation sharing. For example, First Draft developed what is likely the most widely distributed intervention to date: Facebook’s “Tips to Spot False News.”<sup>17</sup> Following the announcement of an “infodemic” by the World Health Organization (WHO) in 2020, First Draft developed text message courses as a low-cost intervention for governments and nongovernmental organizations to distribute. In addition to being relatively low-cost to develop and distribute, the text message course delivery mechanism makes the content more accessible to users in lower-income countries than online games and videos, which can be costly to engage with because they use data on mobile phones. We collaborated with First Draft to develop three distinct

---

<sup>14</sup>Kenya was selected by the Africa Infodemic Response Alliance in the World Health Organization due to the prevalence of misinformation on Kenyan social media.

<sup>15</sup>We paid KSH 55 to participants who did not correctly answer all attention checks to compensate them for their time, but participants were not told in advance we would do so.

<sup>16</sup>87% of participants chose to receive the course via WhatsApp.

<sup>17</sup>Guess et al. (2020) find that these tips, which were promoted at the top of Facebook users’ news feed in 14 countries in April 2017, improved misinformation discernment in a survey experiment similar to ours.

versions of this text message course to test two approaches to decreasing misinformation sharing and their interaction.

First, the Emotions course is motivated by the idea that people may not stop to think about whether a post is misinformation before sharing it because they make an emotions-driven decision. The course focuses on teaching a “Stop and Question” strategy to counteract emotion-based techniques used in social media posts to exploit people’s intuitive reaction to emotional stimulation. In particular, the course teaches that evoking emotions like fear, anger, and superiority are techniques often used to induce users to make an emotions-driven sharing decision.

Second, the Reasoning course is motivated by the idea that people may share misinformation because they *do not know* a post contains misinformation. The course focuses on teaching reasoning-based techniques used in social media posts to exploit a user’s lack of knowledge about context, sources, and other information, then teaching strategies to evaluate whether a post contains misinformation. Specific techniques taught by the course concern misleading graphs, imposter websites, and eyewitness media.

Each of these courses target a different point in the decision-making process in which a user first decides whether to evaluate a post or share it immediately; conditional on not sharing in this first stage, the user employs the strategies at their disposal to evaluate the post and then make a sharing decision. The Emotions course focuses on inducing people to pause in the first stage to evaluate whether a post is misinformation, which is a necessary precondition for the user to apply the strategies learned in the Reasoning course to evaluate the post. On the other hand, pausing to evaluate whether a post is misinformation is irrelevant if the user is unable to evaluate the post and distinguish non-misinformation from misinformation. In other words, the strategies taught in both courses may be necessary to have the greatest impact on misinformation sharing.

The third Combo course teaches the concepts from both the Emotions course and the Reasoning course to test if these approaches are complementary. To ensure that both the length of the text messages and the number of days in the course are the same as the Emotions and Reasoning courses, the content of those courses was streamlined.

In addition to the three treatment courses, there were two baselines. In the No-course baseline, participants received no text messages between the pre-survey and the post-survey. Conditional on completing the post-course survey, participants in the No-course baseline received the Combo course.<sup>18</sup> The Facts baseline sent a text message each day with a fact about misinformation, such as “During the Roman Empire...leaders used misinformation to come to power.” The purpose of the Facts baseline was to separately identify the effect of the educational content in the course and the potential salience effect from a daily text message on misinformation. All courses are provided in Online Appendix B.

---

<sup>18</sup>As participants in the No-course baseline were paid upon completing the post-survey, they received the course but were not paid to engage with it.

## 2.1.2 Survey Structure

This section details the pre-survey, which participants accessed by clicking on a Facebook ad that linked them to the survey on Qualtrics, and the post-survey, which participants accessed by clicking on a link in the last text message they received as a part of the text message course intervention. Figure 2 shows the order of questions for each of the pre- and post-surveys. In the pre-survey, participants responded to social media posts, gave reasons for sharing specific posts, provided demographic information, and enrolled in a “free text message course that teaches you how to protect against misinformation.” In the post-survey, participants only responded to social media posts and again gave reasons for sharing specific posts. Online Appendix C contains screenshots of the pre-survey.<sup>19</sup>



Figure 2: Pre- and Post-survey Structure

Both the pre-survey and the post-survey showed participants ten social media posts and asked them two questions for each post:

1. Would you share this post? (Binary yes/no)
2. To the best of your knowledge, how accurate is the claim in the above post? (4-point Likert scale)

The accuracy nudge randomization that determined the order in which those questions were asked is detailed towards the end of Subsection 2.1.2.

<sup>19</sup>The post-survey asked the sharing and accuracy questions in the same manner as the pre-survey, so screenshots of the post-survey provide no new information.

At the end of the 20-question series (ten posts times two questions), we asked two open-ended free text questions to collect information about participants' reported reasons for their sharing decisions. Participants were shown a post that they previously reported wanting to (not) share, then asked "What made you (NOT) want to share it?" If they did not intend to share any of the posts, they did not see the free-text question on sharing. Similarly, if they always reported wanting to share, they did not see the free-text question on not sharing.

Participants saw one social media post used as an attention check and nine other posts, which were drawn from 62 social media posts we created for the pre- and post-surveys.<sup>20</sup> To create 60 of the posts, we started from 15 health-related facts, verified by a fact-checking organization or published in an academic paper. For each fact, we created four posts. In addition, we created two attention check posts, randomizing which post participants saw in the pre- versus the post-survey. The four posts we created based on each fact were a:

- Baseline non-misinformation post ("non-misinformation")
- Misinformation post using an emotion-based technique ("Emotions")
- Misinformation post using a reasoning-based technique ("Reasoning")
- Misinformation post using both emotion- and reasoning-based techniques ("Combo")

We developed these four versions so that we could observe a participant's behavior when the same information was presented with and without manipulation. The correspondence of the types of misinformation to the treatment courses allows us to learn whether the treatment courses change behavior only on posts that employ the techniques taught therein or whether they have cross-technique effectiveness. If the Emotions course only changes behavior on Emotions posts, and similarly the Reasoning course only changes behavior on Reasoning posts, the policy implications would differ than if one course is as good or better on all types of misinformation (as we find).

In both the pre- and post-survey, each participant saw three non-misinformation posts, two Emotions posts, two Reasoning posts, two Combo posts, and one attention check post, in random order, for a total of ten posts. Importantly, the three non-misinformation posts a participant saw in the pre-survey were matched based on the "fact" to one Emotions, one Reasoning, and one Combo misinformation post the participant saw in the post-survey. We use these three non-misinformation posts in the pre-survey and their corresponding misinformation posts in the post-survey to construct our outcome. Each of the *other* posts shown to participants in the pre- and post-survey were randomly drawn from one of the remaining 12 facts, without replacement. Subsection 2.1.2 details the development of these posts.

The idea for this outcome measure is that the non-misinformation posts allow us to observe a participant's baseline propensity to share a particular fact in the pre-survey. Then, we

---

<sup>20</sup>We created an additional two non-misinformation posts and two attention check posts for the follow-up survey.

Fact	Pre	Post	Corr/Non-corr
1	non-misinformation	Emotions	Corr
2	non-misinformation	Reasoning	Corr
3	non-misinformation	Combo	Corr
4	None	Emotions	Non-corr
5	None	Reasoning	Non-corr
6	None	Combo	Non-corr
7	None	non-misinformation	Non-corr
8	None	non-misinformation	Non-corr
9	None	non-misinformation	Non-corr
10	Emotions	None	Non-corr
11	Emotions	None	Non-corr
12	Reasoning	None	Non-corr
13	Reasoning	None	Non-corr
14	Combo	None	Non-corr
15	Combo	None	Non-corr

Table 1: Example of Posts for each Fact, by Pre-Survey and Post-Survey

*Notes:* The first column indicates the fact number. The second and third columns indicate an example random draw of posts to show a participants in the pre-survey and post-survey, respectively. The fourth column shows for which facts the example participant sees a non-misinformation post in the pre-survey and a corresponding misinformation post in the post-survey about the same fact.

can see how each intervention affects the participant’s propensity to share the same fact presented as misinformation (i.e., we are controlling for a participant’s interest in sharing posts about a specific topic). So, when we observe that a participant reports intending to not share a misinformation post about autism in the post-survey, for example, our primary outcome measure rules out the possibility that the participant is not interested in posts about autism generally. We also test alternative outcome measures in Results section 3.2 and find that our results are robust.

**Social Media Posts** The posts participants see in the surveys were designed to have the same information in all four versions of a post based on one fact. We kept the social media poster and their profile picture the same for all versions of a post, although these elements changed from fact to fact. We varied other characteristics between posts and across facts. Specifically, we varied the length and tone of the posts, and the presence or absence of an image, link, and/or hashtags on the posts. The heterogeneity in the posts was designed to make the detection of misinformation nontrivial. For example, we did not want participants to note that exclamation points always appeared in misinformation but not in non-misinformation posts. While exclamation points may be correlated with misinformation, they do not make a post misinformation.

The three types of misinformation posts used techniques taught in the three treatment courses described in section 2.1.1, so that we could learn whether the treatment courses changed behavior only on posts that employed the techniques taught therein or had cross-technique effectiveness. Emotions posts used emotional and moral-laden language. We targeted emotions taught in the course (fear, anger, and superiority), as well as other emotions such as making people feel like something *must* be wrong. Reasoning posts used techniques from the course and from the FLICC taxonomy (Fake expert, Logical fallacies, Impossible expectations, Cherry picking, Conspiracy theories) developed by Cook (2020). We targeted concepts not taught in the course, as well as those taught in the course, to avoid “teaching to the test.” Online Appendix D contains examples of the posts used in the survey.

We aimed to make the posts realistic representations of what social media users would see in Kenya. Posters were designed to be media reporters and other individuals who would not be in a user’s network. Social media users would see these types of posts through groups on Facebook or public forums such as Twitter. To verify that the posts were representative of what Kenyans see on the internet, we recruited Kenyan workers on Upwork to evaluate the content of our posts, and we made changes based on their suggestions.

**Accuracy Nudge** Participants were randomly assigned to see the accuracy nudge at the beginning of the pre-survey. This treatment assignment determined the *order* of the first 20 questions, which elicited sharing intentions and accuracy for each of ten posts in both the pre- and post-survey. Figure 2 shows the question order for those who were, and were not, assigned to see the accuracy nudge. Participants treated with the accuracy nudge (Accuracy Inter group) were first asked about the accuracy of a post and then were asked to make a sharing decision about the same post, for each of the ten posts (i.e., questions interweave). Participants not treated with the accuracy nudge (Accuracy After group) were asked to make a sharing decision about all ten posts and then were asked about the accuracy of all ten posts.

Accuracy nudges have been shown to be an effective means of decreasing misinformation sharing in laboratory experiments, survey experiments, and on Twitter (Offer-Westort et al., 2021; Pennycook et al., 2021). The mechanism through which accuracy nudges aim to treat misinformation sharing – priming a person to think about the accuracy of the post – has synergies with strategies in both the Emotions and Reasoning courses. The accuracy nudge forces users to pause before making the sharing decision, a strategy the Emotions course encourages users to employ. The accuracy nudge also encourages users to evaluate the accuracy of posts before making a sharing decision, for which the Reasoning course provides strategies.

This survey design element allows us to achieve two goals. First, we can learn whether the text message course provides educational value beyond encouraging users to pause and evaluate posts before sharing them. For example, in the Emotions course, the “Stop and Question” strategy is motivated by describing how emotional language is used to mislead users and manipulate them into sharing. This educational content could increase participants’ desire to avoid sharing misinformation, and so the accuracy nudge allows us to evaluate whether that content has added value. Second, as an externally-validated intervention, the accuracy nudge

provides an internally-valid benchmark for the text message courses.

## 2.2 Stage 2: Follow-up Experiment

The experimental design to this point was prespecified. When we learned that the treatment effect sizes for short-run outcomes were sufficiently high that we were powered to detect reasonable long-run effects, we executed a follow-up survey. We recruited participants for the follow-up survey by sending a text message informing them that they could earn an additional KSH 350 by completing another survey.

Before sending the text, participants were randomized on whether the recruitment text reminded participants about course content through a “priming” treatment. These primes were customized to the course to which the participant was exposed in the Main Experiment to remind participants about what they previously learned. Figure 3 shows an example of the text message a participant in the Emotions course received with the prime. Online Appendix E contains the text for all primes.



😬 Have you noticed any posts recently trying to create feelings of fear, anger, or superiority?! Just a quick reminder to STOP and QUESTION the information in the post when that happens. You don't want to share anything you're not 100% sure is true!

Want to earn 🇰🇸 KSH 350 🇰🇸 in mobile airtime TODAY? Take our 10-question survey now by replying START.

Figure 3: Message Sent to Emotions Course Participants Randomized into the Prime

### 2.2.1 Survey Structure

The follow-up survey showed participants five social media posts and asked them the same sharing question as in the pre- and post-survey. We did not ask the accuracy question (and so we did not implement the accuracy nudge) in the follow-up survey. Then, we asked five reflective questions:

1. What are some techniques that people use to create misleading social media posts?
2. When browsing your time line in the last month, did you notice any post that looked misleading? If so, what made it seem misleading?
3. How did you feel when you saw the misleading post? If you haven't seen any misleading posts recently, how do you think you would feel?
4. Has the Inoculation against Misinformation course changed your behavior on social media? If so, how?



5. If you were to tell a friend what you learned in the course, what tip would you share?

The five posts that participants saw were a non-misinformation post, an Emotions post, an Reasoning post, a Combo post, and an attention check post. Similar to how the post-survey matched misinformation posts to non-misinformation posts in the pre-survey, the follow-up matched misinformation posts to non-misinformation posts a participant saw in the post-survey.

## 2.3 Implementation

The Stage 1: Main Experiment was implemented in July 2022, with all participants completing the text message course intervention and post-survey by August 2022. The Stage 2: Follow-up Experiment was launched and concluded in September 2022. Table 2 shows the funnel of participants from starting the pre-survey to completing the follow-up survey. Attrition overall was fairly low due to the monetary incentive for completion, with 34% of participants who engaged with us completing the Main Experiment, and 61% of participants invited to the Follow-up Experiment (based on completing post-survey) completing it. More than 90% of participants who started each of our surveys (pre-, post-, or follow-up) completed them, suggesting the surveys were not too burdensome. Attrition during the text message course was higher, with only 66% of those who started the course completing it, but this attrition rate is still quite low considering we were implementing a five-day intervention.

Furthermore, Table 3 shows that attrition was not substantially different between assignment groups in the Main Experiment, with 36-43% of participants who completed the pre-survey completing the post-survey in all groups. Attrition was also not substantially different between assignment groups in the Follow-up Experiment, with 59-63% of participants who completed the post-survey also completing the follow-up survey in all groups. We show little difference in the proportion of participants who correctly answered attention check questions across intervention assignment groups; however, we do see sizeable differences between the Accuracy Inter and Accuracy After assignment groups. Participants in the Accuracy Inter group are consistently six to nine percentage points more likely to correctly answer both attention check questions, suggesting that the Accuracy Inter participants were paying closer attention. In Appendix D, we show that our results are robust to restricting our sample to those who correctly answer the attention check questions.

	N Participants	% of "Started Pre-survey"	% of Previous Funnel Stage
<b>Started Pre-survey</b>	25,287	-	-
<b>Completed Pre-survey</b>	22,526	89.08%	89.08%
<b>Started Text Message Course</b>	18,598	73.55%	82.56%
<b>Completed Day 1 Course</b>	16,684	65.98%	89.71%
<b>Completed Day 2 Course</b>	13,997	55.35%	83.89%
<b>Completed Day 3 Course</b>	13,093	51.78%	93.54%
<b>Completed Day 4 Course</b>	11,396	45.07%	87.04%
<b>Completed Entire Course</b>	10,934	43.24%	95.95%
<b>Started Post-survey</b>	9,589	37.92%	87.70%
<b>Completed Post-survey</b>	8,684	34.34%	90.56%
<b>Started Follow-up</b>	5,785	22.88%	66.62%
<b>Completed Follow-up</b>	5,316	21.02%	91.89%

Table 2: Funnel of Participants

*Notes:* The number of participants who started the pre-survey is an upper bound estimate because we could count only survey copies on Qualtrics and not identify users. We were only able to identify users once they completed the pre-survey and provided a phone number, which is how we ensured that all down-funnel outcomes counted unique users based on phone number. The post- and follow-up surveys required users to validate their phone number before they could start the survey. We received 40,845 survey copies in total for the pre-survey, of which we discarded 3,092 users who had participated in one of our pilot studies and another 12,466 survey copies filled out by duplicated phone numbers. In the post-survey, we filtered out participants who encountered system errors (N = 104), did not have at least 5 days in between pre- and post-survey dates (N = 2,101), and/or did not complete the full text-message course (N = 509). For the follow-up survey, we filtered out 2,369 survey copies filled out by duplicated phone numbers and retained only the first copy completed by each phone number.

	No-course baseline	Facts baseline	Reasoning course	Emotions course	Combo course	Totals
<b>Phase 1: Main Exp.</b>						
Accuracy After	827 (36.46%) [30.96%]	886 (39.05%) [32.96%]	856 (37.93%) [31.89%]	894 (38.67%) [33.45%]	859 (37.82%) [34.92%]	4,322 (37.99%) [32.86%]
Accuracy Inter	801 (36.05%) [37.45%]	904 (39.95%) [41.26%]	834 (37.72%) [40.89%]	942 (42.72%) [40.76%]	881 (39.19%) [40.64%]	4,362 (39.12%) [40.26%]
<b>Totals</b>	1,628 (36.26%) [34.15%]	1,790 (39.50%) [37.15%]	1,690 (37.82%) [36.33%]	1,836 (40.65%) [37.20%]	1,740 (38.50%) [37.82%]	8,684 (38.55%) [36.57%]
<b>Phase 2: Follow-up Exp.</b>						
	964 (59.21%) [85.48%] {43.36%}	1,069 (59.72%) [86.25%] {44.53%}	1,072 (63.43%) [86.85%] {41.14%}	1,122 (61.11%) [84.05%] {42.60%}	1,089 (62.59%) [85.86%] {43.80%}	5,316 (61.22%) [85.68%] {43.08%}

Table 3: Attrition and Attention Check Passing Rates, by Intervention Assignment Group

*Notes:* Main cells show the number of participants who completed the post-survey (in the Main Experiment) and the number of participants who completed the follow-up (in the Follow-up Experiment) in their respective assignment group. Numbers in parentheses are the ratios of the number of participants who completed the post-survey to the number of participants who completed the pre-survey (in the Main Experiment) and the ratios of the number of participants who completed the follow-up survey to the number of participants who completed the post-survey (in the Follow-up Experiment). Numbers in brackets are the ratios of the number of participants who correctly answered all attention checks in the pre- and post-survey to the number of participants who completed the post-survey (in the Main Experiment) and the ratios of the number of participants who correctly answered the single attention check in the follow-up survey to the number of participants who completed the follow-up survey (in the Follow-up Experiment). Numbers in curly brackets are the ratios of the number of participants who correctly answered all attention check questions in the pre-, post-, and follow-up surveys to the number of participants who completed the follow-up survey.

The descriptive statistics in Table 4 show that our sample is young at approximately 26 years old, disproportionately men, mostly single, and predominantly Christian. Participants are more likely to be unemployed than employed (part-time, full-time, or self-employed), and they are about evenly split between rural, urban, and suburban communities.

	Mean	Std. Dev	Q1	Median	Q3	% Missing
<b>Age</b>	26.397	7.5918	22	25	29	0%
<b>Man</b>	65.4%	0.4757				0%
<b>Educational Attainment</b>						
High school or less	27.1%	0.444				0%
Some college	35.5%	0.4787				0%
Bachelor's degree	36.1%	0.4802				0%
Graduate degree	1.3%	0.1118				0%
<b>Married</b>	32.6%	0.4686				0%
<b>Employment Status</b>						
Unemployed	40.1%	0.4901				0%
Employed	32.5%	0.4684				0%
Student	27.4%	0.4459				0%
<b>Location</b>						
Mostly urban	28.5%	0.4514				0%
Suburban	38.5%	0.4867				0%
Mostly rural	33.0%	0.4701				0%
<b>Christian</b>	94.2%	0.2340				0%
<b>Attends religious services</b>	94.6%	0.2256				0%
<b>Uses social media</b>	99.33%	0.0815				0%
<b>Hrs/day on social media</b>	5.779	3.7131	3	5	7	0%
<b>Prop. of content shared</b>						
0-20%	17.6%	0.3811				0%
20-40%	20.0%	0.4000				0%
40-60%	30.6%	0.4610				0%
60-80%	23.3%	0.4227				0%
80-100%	8.4%	0.2780				0%
<b>Pre-survey Sharing</b>						
All posts	0.619	0.3260	0.444	0.667	0.889	0%
Non-misinformation posts	0.682	0.3616	0.333	0.667	1	0%
Misinformation posts	0.588	0.3464	0.333	0.667	1	0%

Table 4: Summary Statistics for Post-Survey Completers Sample

*Notes:* Sample includes the 8,684 participants who completed the post-survey. See Table C3 in the Appendix for definitions of covariates.

Appendix C.3 contains figures that show attrition and balance. We first consider what kind of selection on observables we observe based on attrition in the post- and follow-up surveys. We observe no differences larger than a tenth of a standard deviation between the participants who completed the post-survey and those who completed the pre-survey only. We observe a few differences between those who completed the follow-up survey and those who completed the post-survey only. Follow-up completers are younger, less-educated, and more likely to be unmarried and to be a student. Then, we consider the balance across the text message course intervention assignment groups. Figure C2 shows that there were no differences greater than a tenth of a standard deviation between assignment groups among participants who completed the post-survey. Importantly, we observed no differences in sharing behavior in the pre-survey.

We do, however, see two small differences among participants who complete the follow-up survey. Participants in the Reasoning course group are less likely to be married and more likely to share non-misinformation posts relative to participants in the Facts baseline. The latter difference could be concerning as it determines the denominator in the primary outcome of interest; therefore, we include Augmented Inverse Propensity Weighted (AIPW) estimates for our results in Appendix F and find that they are robust to controlling for observed covariates.

### 3 Results

This section evaluates the results of the experiment. Section 3.0.1 defines the outcomes of interest and presents summary statistics for outcomes by treatment assignment status. Section 3.1 analyzes treatment effects for the text message courses. Section 3.2 presents results with alternative outcome measures, while Section 3.3 establishes that the results persist in the long run. Section 3.4 analyzes the accuracy nudge. In Appendix A, we present all of the results specified in our pre-analysis plan (Athey et al., 2022).

#### 3.0.1 Defining and Summarizing Outcomes

To define the primary outcome, let  $S_i^{N,pre}(j)$  be participant  $i$ 's sharing decision for the non-misinformation  $N$  post about fact  $j \in \{1, 2, 3\}$  in the pre-survey and  $S_i^{M,post}(j)$  their sharing decision for the misinformation  $M$  post about fact  $j$  in the post-survey. Then, our sharing outcome is constructed as follows:

$$share_i = \frac{\sum_{j=1}^3 1 \left[ S_i^{M,post}(j) | S_i^{N,pre}(j) = 1 \right]}{\sum_{j=1}^3 S_i^{N,pre}(j)} \quad (1)$$

The denominator is the number of non-misinformation posts participant  $i$  shares in the pre-survey out of the three they are shown. The numerator is the number of misinformation posts corresponding to the non-misinformation posts participant  $i$  shares in the pre-survey that participant  $i$  shares in the post-survey (recall the example in Table 1). So, a participant who shares all three non-misinformation posts in the pre-survey can share  $\{0, 1, 2, 3\}$  corresponding posts in the post-survey and their outcome can take on the value  $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$ . A participant who shares only two of the non-misinformation posts in the pre-survey can share  $\{0, 1, 2\}$  corresponding posts in the post-survey, where the observation of the misinformation post corresponding to the non-misinformation post that the participant does not share in the pre-survey is dropped, and their outcome can take on the value  $\{0, \frac{1}{2}, 1\}$ . 994 participants who share zero of the non-misinformation posts in the pre-survey are dropped from the analysis for this outcome. Because sharing in the pre-survey differs between accuracy nudge groups, we differentially drop participants in the Accuracy Inter (656) and Accuracy

After groups (338). This differential dropping is not problematic because accuracy nudge groups are randomized into text message course assignment groups.

This outcome is designed to disentangle the part of the sharing decision that is related to the participant’s evaluation of a post as misinformation from the part that is related to the participant’s interest in the post’s content. By measuring baseline preferences for post content with no misinformation, we can isolate the part of the sharing decision that is based on the participant’s evaluation of a post as misinformation alone. Subsection 3.2 shows that participants are 43.8% less likely to share misinformation posts corresponding to non-misinformation posts they do *not* intend to share in the pre-survey than misinformation posts corresponding to non-misinformation posts they do intend to share (22.5% compared to 51.5%, on average), indicating that participant interest in content is an important driver of variation in sharing overall. Thus, accounting for interest can be helpful for reducing variation in our misinformation sharing outcome measure.

In addition to our primary outcome, we also consider outcomes that relate to the sharing of posts that are not misinformation. If our treatments affect sharing of all posts equally, then they cannot be interpreted as specifically targeting misinformation. An outcome commonly used in the literature is sharing *discernment*, which is the difference between the number of non-misinformation posts shared and the number of misinformation posts shared.<sup>21</sup> Another discernment outcome we consider takes the difference in the accuracy score of non-misinformation posts and the accuracy score of the misinformation posts. Section 3.2 examines these and additional outcome measures that do not include the posts used to construct the primary outcome.

Table 5 presents means and standard errors for a variety of outcomes, broken out by both the text message course assignment group and assignment to the accuracy nudge. Outcomes include Sharing Rate (Equation 1), discernment, and the proportion of misinformation and non-misinformation posts shared in the pre-survey and the post-survey. Table 5 forms the basis of the treatment effect estimates presented in subsequent sections.

### 3.1 Text Message Treatment Course Effectiveness

We answer the following two questions to evaluate the text message treatment courses’ effectiveness. Do the treatment course have different outcomes than our baselines? And do outcomes differ across the Emotions, Reasoning, and Combo treatment courses? Figure 4 illustrates averages for the Sharing Rate outcome across each of the text message course interventions for the 7,774 participants who both completed the post-survey and shared at least one non-misinformation post in the pre-survey.

Several issues inform the interpretation of the results. First, note that our sample is selected based on completing the post-survey, and so treatment effects should be interpreted in that context. However, as described in Section 2.3, selected individuals have similar observable

---

<sup>21</sup>Offer-Westort et al. (2021) studies a variant where true posts are given weight .5 relative to false posts receiving weight 1.

	No-course baseline	Facts baseline	Reasoning course	Emotions course	Combo course	Treatment courses	Baselines	All
<b>Number of Observations</b>								
Accuracy After	827	886	856	894	859	2,609	1,713	4,322
Accuracy Inter	801	904	834	942	881	2,657	1,705	4,362
All	1,628	1,790	1,690	1,836	1,740	5,266	3,418	8,684
<b>Pre Non-misinfo Posts</b>								
Accuracy After	0.650	0.628	0.654	0.642	0.648	0.648	0.638	0.644
	(0.0129)	(0.0118)	(0.0124)	(0.0117)	(0.0124)	(0.0070)	(0.0087)	(0.0055)
Accuracy Inter	0.756	0.739	0.752	0.738	0.726	0.739	0.747	0.742
	(0.0105)	(0.0107)	(0.0108)	(0.0109)	(0.0111)	(0.0063)	(0.0075)	(0.0048)
All	0.704	0.683	0.704	0.689	0.686	0.693	0.693	0.693
	(0.0084)	(0.0081)	(0.0083)	(0.0081)	(0.0084)	(0.0048)	(0.0058)	(0.0037)
<b>Pre Misinfo Posts</b>								
Accuracy After	0.525	0.497	0.503	0.519	0.519	0.514	0.510	0.512
	(0.0125)	(0.0113)	(0.0120)	(0.0113)	(0.0117)	(0.0067)	(0.0084)	(0.0052)
Accuracy Inter	0.604	0.581	0.576	0.563	0.574	0.571	0.592	0.580
	(0.0104)	(0.0103)	(0.0106)	(0.0107)	(0.0107)	(0.0062)	(0.0073)	(0.0047)
All	0.565	0.539	0.540	0.541	0.546	0.542	0.551	0.546
	(0.0081)	(0.0077)	(0.0081)	(0.0078)	(0.0080)	(0.0046)	(0.0056)	(0.0035)
<b>Post Non-misinfo Posts</b>								
Accuracy After	0.639	0.558	0.518	0.477	0.486	0.493	0.596	0.533
	(0.0131)	(0.0129)	(0.0131)	(0.0122)	(0.0128)	(0.0073)	(0.0092)	(0.0058)
Accuracy Inter	0.719	0.657	0.582	0.563	0.559	0.568	0.687	0.615
	(0.0116)	(0.0122)	(0.0125)	(0.0123)	(0.0129)	(0.0073)	(0.0085)	(0.0056)
All	0.680	0.607	0.550	0.519	0.522	0.530	0.642	0.574
	(0.0088)	(0.0090)	(0.0091)	(0.0087)	(0.0091)	(0.0052)	(0.0063)	(0.0040)
<b>Post Misinfo Posts</b>								
Accuracy After	0.513	0.427	0.380	0.326	0.356	0.353	0.468	0.398
	(0.0122)	(0.0118)	(0.0116)	(0.0103)	(0.0116)	(0.0064)	(0.0085)	(0.0052)
Accuracy Inter	0.578	0.504	0.427	0.381	0.387	0.398	0.539	0.454
	(0.0111)	(0.0116)	(0.0114)	(0.0106)	(0.0112)	(0.0064)	(0.0081)	(0.0051)
All	0.546	0.465	0.404	0.353	0.371	0.375	0.504	0.426
	(0.0083)	(0.0083)	(0.0082)	(0.0074)	(0.0081)	(0.0045)	(0.0059)	(0.0037)
<b>Primary Outcome: Sharing Rate</b>								
Accuracy After	0.631	0.544	0.495	0.413	0.461	0.455	0.584	0.505
	(0.0151)	(0.0151)	(0.0150)	(0.0140)	(0.0154)	(0.0085)	(0.0107)	(0.0068)
Accuracy Inter	0.647	0.582	0.498	0.447	0.447	0.464	0.614	0.524
	(0.0134)	(0.0140)	(0.0143)	(0.0138)	(0.0142)	(0.0081)	(0.0097)	(0.0064)
All	0.639	0.563	0.496	0.430	0.454	0.459	0.600	0.515
	(0.0100)	(0.0103)	(0.0103)	(0.0098)	(0.0105)	(0.0059)	(0.0072)	(0.0046)
<b>Pre Sharing Discernment Score</b>								
Accuracy After	-1.196	-1.100	-1.054	-1.188	-1.169	-1.140	-1.145	-1.142
	(0.0581)	(0.0525)	(0.0565)	(0.0526)	(0.0520)	(0.0310)	(0.0390)	(0.0243)
Accuracy Inter	-1.357	-1.269	-1.201	-1.167	-1.268	-1.211	-1.311	-1.251
	(0.0531)	(0.0513)	(0.0534)	(0.0547)	(0.0548)	(0.0314)	(0.0369)	(0.0239)
All	-1.278	-1.183	-1.128	-1.178	-1.218	-1.175	-1.228	-1.196
	(0.0393)	(0.0367)	(0.0389)	(0.0379)	(0.0377)	(0.0220)	(0.0269)	(0.0170)
<b>Post Sharing Discernment Score</b>								
Accuracy After	-1.164	-0.889	-0.728	-0.525	-0.678	-0.639	-1.018	-0.787
	(0.0555)	(0.0529)	(0.0529)	(0.0470)	(0.0519)	(0.0292)	(0.0384)	(0.0234)
Accuracy Inter	-1.308	-1.050	-0.818	-0.597	-0.646	-0.686	-1.175	-0.879
	(0.0531)	(0.0531)	(0.0556)	(0.0516)	(0.0533)	(0.0309)	(0.0377)	(0.0242)
All	-1.237	-0.969	-0.773	-0.560	-0.662	-0.662	-1.097	-0.833
	(0.0384)	(0.0375)	(0.0384)	(0.0348)	(0.0371)	(0.0212)	(0.0269)	(0.0168)
<b>Pre Accuracy Discernment Score</b>								
Accuracy After	1.924	2.394	2.803	2.176	2.283	2.408	2.173	2.316
	(0.2555)	(0.2323)	(0.2489)	(0.2271)	(0.2373)	(0.1371)	(0.1720)	(0.1072)
Accuracy Inter	1.232	1.445	1.451	1.843	1.799	1.700	1.342	1.558
	(0.2402)	(0.2231)	(0.2315)	(0.2348)	(0.2373)	(0.1354)	(0.1635)	(0.1044)
All	1.572	1.924	2.118	2.014	2.044	2.057	1.757	1.939
	(0.1753)	(0.1614)	(0.1706)	(0.1633)	(0.1678)	(0.0965)	(0.1189)	(0.0749)
<b>Post Accuracy Discernment Score</b>								
Accuracy After	1.779	3.319	3.645	4.142	3.872	3.896	2.595	3.388
	(0.2362)	(0.2326)	(0.2292)	(0.2089)	(0.2273)	(0.1278)	(0.1669)	(0.1020)
Accuracy Inter	1.063	2.280	2.916	3.347	3.165	3.146	1.692	2.570
	(0.2336)	(0.2226)	(0.2269)	(0.2142)	(0.2232)	(0.1278)	(0.1618)	(0.1009)
All	1.415	2.804	3.276	3.755	3.523	3.524	2.143	2.981
	(0.1663)	(0.1615)	(0.1614)	(0.1498)	(0.1595)	(0.0905)	(0.1164)	(0.0719)

Table 5: Summary of Outcomes by Text Message Course Intervention and Accuracy Nudge Assignment

Notes: Sample includes the 8,684 participants who completed the post-survey, except for the “Primary Outcome: Sharing Rate” rows that exclude 996 participants who did not share at least one non-misinformation post in the pre-survey. The first two rows display the number of observations in each assignment group. The other rows display averages by assignment group, with standard errors in parentheses.

covariates to the overall recruited population, and are comparable across treatment arms.

Second, as we prespecified, other than when we analyze the accuracy nudge specifically in subsection 3.4, we pool the two accuracy nudge groups when analyzing the text message course. The detailed breakdowns for each accuracy nudge condition are included in Table 5. When interpreting pooled results for the Sharing Rate, note that Table 5 shows that the accuracy questions interweaved between the sharing questions in the Accuracy Inter group decrease misinformation sharing relative to the Accuracy After group in both the pre- and post-surveys.<sup>22</sup> This result means that the distribution of our denominator that serves as a baseline in our primary outcome (as well as the numerator) differs between the two accuracy nudge groups. Since assignment to the two accuracy conditions is balanced across the text message course assignments, this issue does not bias our results.

Figure 4 illustrates contrasts between the treatment courses and baselines. Relative to the No-course baseline sharing rate of 63.9 p.p., the treatment courses decrease the Sharing Rate by 18.1 p.p. ( $SE = 1.16$  p.p.) or approximately 28%. Relative to the Facts baseline, the treatment course decreases the Sharing Rate by 10.4 p.p. ( $SE = 1.19$  p.p.) or approximately 18% relative to the 56.3 p.p. sharing rate for the Facts baseline. The difference between the No-course and Facts baselines of 7.7 p.p. ( $SE = 1.44$  p.p.) shows that making misinformation salient to participants on a daily basis does have some beneficial impact on sharing behavior, but the treatment courses are providing educational content beyond this salience effect.

Figure 4 also shows that the best-performing course is the Emotions course. Further, the Emotions course reduces misinformation sharing by 2.3 p.p. ( $SE = 1.44$ ) more than the Combo course. Thus, we can bound any benefit of the Combo course to no more than 0.5 p.p., which is 1.1% of the 45.8 p.p. baseline sharing rate in the treatment courses in the post-survey or 2.8% of the 18.1 p.p. treatment effect of the treatment courses relative to the No-course baseline. This result implies that teaching the reasoning-based techniques in the Combo course is, at best, not harmful.<sup>23</sup> In comparison to the Reasoning course, the Emotions course decreases misinformation sharing by 6.6 p.p. ( $SE = 1.43$  p.p.) more. The Combo course also decreases misinformation sharing more than the Reasoning course by 4.3 p.p. ( $SE = 1.47$ ), further confirming the importance of focusing on emotion-based techniques. In summary, teaching emotion-based techniques decreases misinformation sharing more than teaching reasoning-based techniques, and the reasoning-based content has no added value that we can detect when combined with the emotion-based content.

### 3.1.1 Heterogeneity by Misinformation Post Type

This section explores the effectiveness of the different treatment courses for different types of posts. In the pre- and post-surveys, we show participants multiple types of misinformation, defined in the same way as the treatment courses: Emotions, Reasoning, and Combo. This

---

<sup>22</sup>We discuss the estimates for the differences between the accuracy nudge groups in subsection 3.4.

<sup>23</sup>Note that when we report p-values for hypothesis tests in Appendix A, the hypotheses tested in this paragraph comparing the treatment courses to one another are grouped in Family 2 with tests broken out by the type of post presented in Figure 5 when correcting for multiple hypothesis testing.

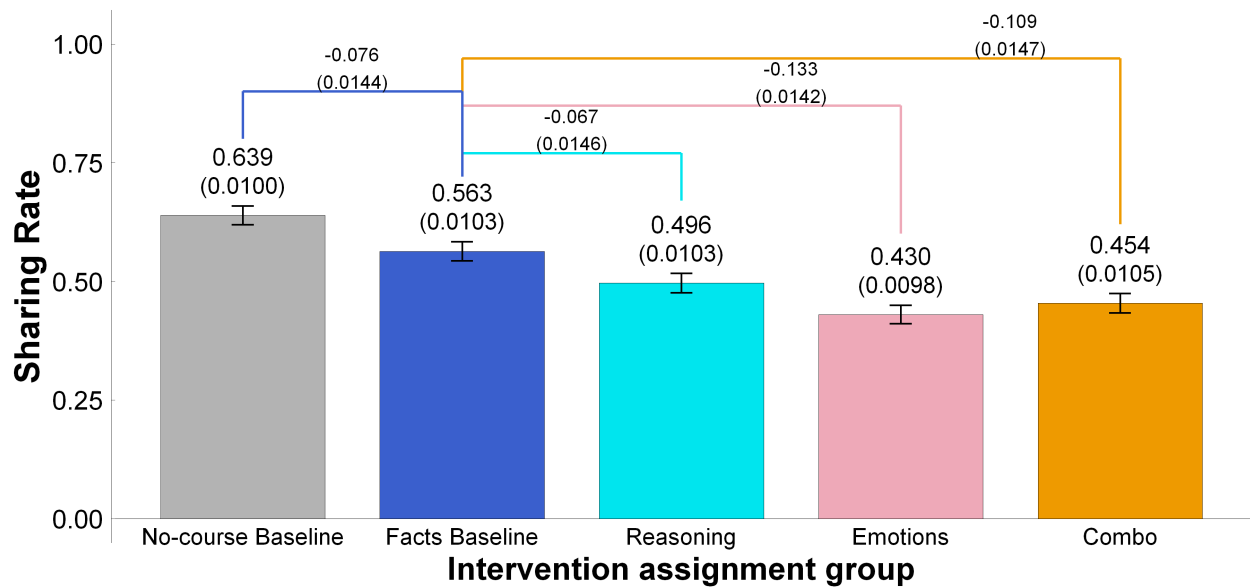


Figure 4: Misinformation Sharing, by Intervention Assignment Group

*Notes:* Sample includes the 7,688 participants who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in Equation 1, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.



correspondence allows us to evaluate whether, for example, participants in the Emotions course are less likely to share Emotions posts, but are no different with respect to Reasoning posts. In particular, based on the result from section 3.1 that the Emotions course outperforms the Reasoning course, we want to learn whether the Reasoning course might outperform the Emotions course in Reasoning posts in particular, even if it performs worse overall. Understanding the cross-technique effect of the treatment courses also contributes to our understanding of how the strategies taught in the courses generalize to techniques not taught.

Figure 5 shows that, not only does the Emotions course outperform the Reasoning and Combo courses in Emotions posts, it does not perform statistically worse on Reasoning or Combo posts. The Emotions course decreases the sharing of Emotions posts by 7.9 p.p. ( $SE = 2.02$  p.p.) more than the Reasoning course and by 5.7 p.p. ( $SE = 2.02$  p.p.) more than the Combo course. Furthermore, the Emotions course decreases misinformation sharing on Combo posts by 5.0 p.p. ( $SE = 2.91$ ) more than the Combo course and by 7.7 p.p. ( $SE = 1.99$  p.p.) more than the Reasoning course. While the Emotions course also decreases misinformation on Reasoning posts by 3.0 p.p. ( $SE = 2.02$  p.p.) more than the Reasoning course, the Combo course outperforms the Emotions course on Reasoning posts by 1.9 p.p. ( $SE = 2.01$  p.p.). We can rule out that the Combo course outperforms the Emotions course by more than 5.9 p.p. on Reasoning posts, and the difference is not statistically distinguishable even with our large sample size ( $p = 0.452$ ). Appendix Table A3 shows the differences for each prespecified pairwise comparison and the Romano-Wolf adjusted p-values. We fail to reject differences between the Emotions course and the Combo course on both the Reasoning and Combo posts as well as the difference between the Emotions course and the Reasoning course on Combo posts.

The result that the Emotions course is best on Emotions posts, and not detectably worse than the other two treatment courses on Reasoning and Combo posts, confirms that the Emotions course is the most effective of the interventions we test. The strategy targeting the first-stage of the decision-making process is changing sharing behavior on all types of misinformation we test, not just the misinformation posts that use the emotion-based techniques taught in the Emotions course.

### 3.1.2 Heterogeneity by Subgroup

We explore the effectiveness of the different treatment courses for different participants' covariates, with the goal of understanding whether certain courses work better for certain subgroups of people. We are primarily interested in the following covariates: gender, age, proportion of content shared on social media, hours on social media, and predicted pre-survey misinfo sharing. We take two approaches in conducting subgroup analysis. First, we consider each covariate individually and look at whether the Emotions or Reasoning course works better for each subgroup of each covariate after adjusting for multiple hypothesis testing (Romano and Wolf, 2007). Second, we take a data-driven approach to estimate Rank-Weighted Average Treatment Effect (RATE) on the Sharing Rate to look for more complex heterogeneity.

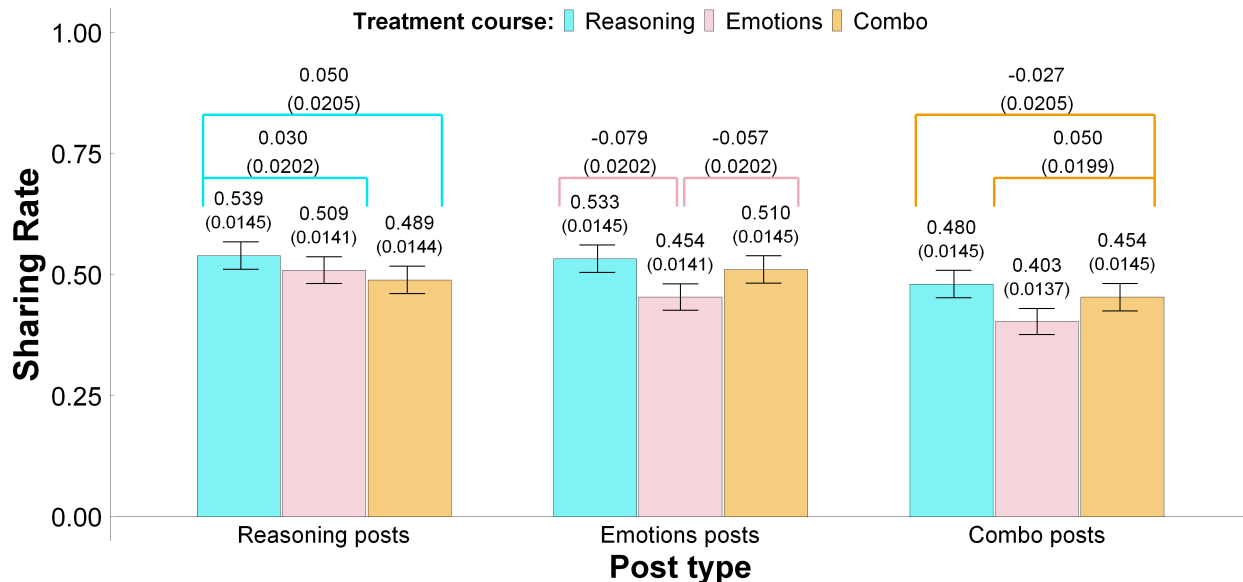


Figure 5: Misinformation Sharing, by Intervention Assignment Group and Post Type

*Notes:* Sample includes 4,646 participants in the Reasoning, Emotions or Combo intervention assignment groups who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Each group of bars displays the Sharing Rate for misinformation posts of each type, as defined in Equation 1, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

When looking at covariates individually, we find differences in the primary outcome only for gender between the Emotions and Reasoning courses. The Emotions course reduces misinformation sharing more for women (14.2 p.p, SE = 1.28 p.p) than for men (2.5 p.p, SE = 1.77 p.p) compared to the Reasoning course. The difference-in-difference estimate—difference between women and men in the difference between the Emotions and Reasoning courses—is large in terms of both levels (11.69 p.p, SE = 2.98 p.p) and percentages (26.2%, SE = 6.53%) (see Appendix Table E1 for full statistics). This result suggests that the Emotions course works better for women participants. We conduct robustness check on the gender differences for the Sharing Rate in the follow-up, the difference in discernment score, and the difference in accuracy discernment, and find similar results (see Appendix Table E2 for the statistics for the three additional outcomes). We further investigate whether such differences are driven by certain post types (see Appendix Figure E3) or facts (see Appendix Figures E4-E5), but we do not find such differences by either specific post types or individual facts.

Our data-driven approach (see Appendix Table E3) finds no heterogeneity between the Emotions and Reasoning courses when we include all covariates (1.1 p.p, SE = 2.11 p.p). However, if we include gender as the only covariate for the data-driven approach, we find that the Emotions course reduces sharing by 3.6 p.p (SE = 1.15 p.p), confirming the result we found in the first approach.

## 3.2 Alternative Outcome Measures

We test four alternative outcome measures in this section to assess the robustness of our results. The first three alternative outcomes vary how we measure misinformation sharing. First, we test one of the standard outcome measures used in the literature: sharing discernment. Second, we evaluate misinformation sharing using only misinformation posts not used in the main analysis. Third, we evaluate the “opposite” of our primary outcome in Equation 1 by sub-setting to non-misinformation posts that participants do NOT intend to share in the pre-survey (instead of sub-setting to the posts participants intend to share). The fourth alternative outcome uses the accuracy scores we elicit in addition to sharing decisions.

The first alternative outcome we assess is sharing discernment, which is computed by taking the difference between the number of non-misinformation posts and the number of misinformation posts a participant intends to share. This outcome is commonly used in the literature to handle concerns that interventions to decrease misinformation sharing may reduce sharing overall, not just sharing of misinformation. Positive values mean the participant is sharing more non-misinformation than misinformation posts, so taking the post- minus pre-survey difference means that more positive values represent *better* discernment.

Figure 6 shows that all treatment courses improve discernment relative to both the No-course baseline and the Facts baseline, with Emotions performing the best of the three treatment courses. Relative to a baseline discernment of -1.28 ( $SE = 0.055$ ), the Emotions course increases discernment by 0.40 more than the Facts baseline. As expected, participants in the No-course baseline show no change in discernment.

The second alternative outcome is the proportion of misinformation posts shared, excluding the three misinformation posts in the post-survey that correspond to non-misinformation posts in the pre-survey. In the pre-survey, the denominator for this measure is six; while for the post-survey, the denominator is only three because three of the six misinformation posts participants saw in the post-survey are used in the primary outcome. The concern we address with this outcome is that seeing the same fact twice (as non-misinformation, then as misinformation) could affect our results through memory, although we posit that the five days in between negate that concern. By using only misinformation posts for which participants saw the related fact just once, we can assess the robustness of our results to this concern. We compare the post- minus pre-survey difference, so negative values mean sharing *less* misinformation in the post-survey.

Figure 7 shows that, again, all treatment courses do better than both baselines at decreasing misinformation sharing. The Emotions course has the largest treatment effect, but it is only 1.6 p.p. ( $SE = 1.26$  p.p.) larger than the treatment effect for the Combo course. Note that there is some decrease in misinformation sharing in the No-course baseline of 2.7 p.p. ( $SE = 0.80$  p.p.), indicating a small change in behavior between the pre- and post-survey even without exposure to a text message course. One possible explanation is that the accuracy questions at the end of the pre-survey affect sharing behavior in the post-survey. The decrease in misinformation sharing from the pre- to the post-survey in the Accuracy After group is 3.8 p.p. ( $SE = 1.23$  p.p.) compared to 1.5 p.p. ( $SE = 1.32$  p.p.) in the

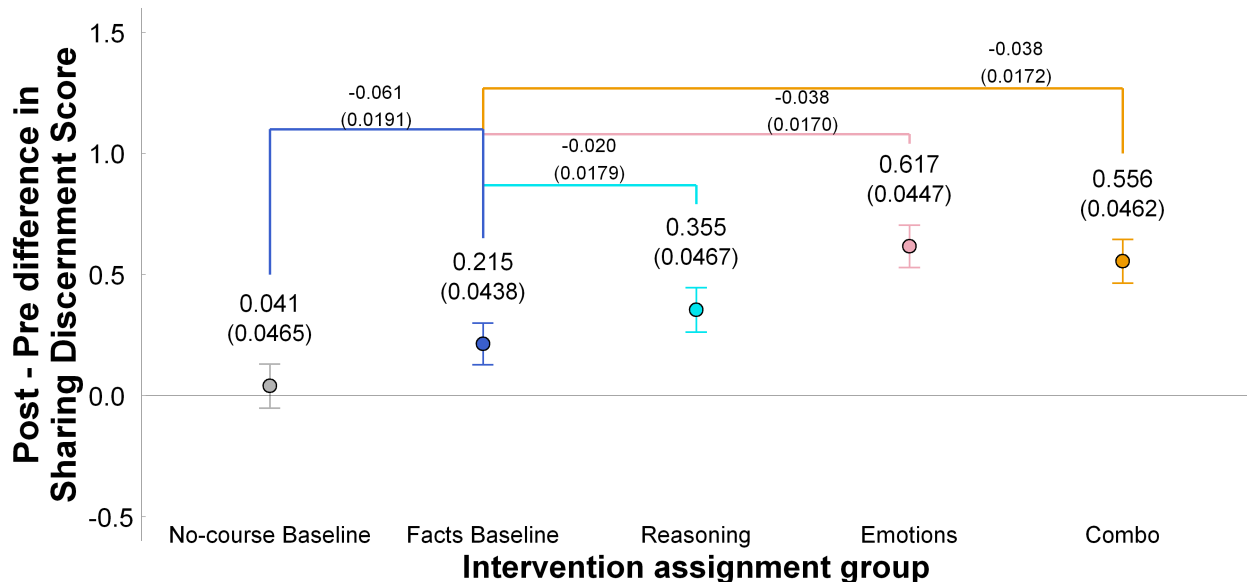


Figure 6: Differences in Sharing Discernment Score, by Intervention Assignment Group

*Notes:* Sample includes the 8,684 participants who completed the post-survey. Each point displays the average post- minus pre-survey difference in sharing discernment score, defined as the number of non-misinformation posts shared minus the number of misinformation posts shared, for participants in their respective intervention assignment group, pooling participants in the Accuracy After and Accuracy Inter groups. Above each point, the standard error of each difference is shown in parentheses below the difference. The thin colored bars represent 95% confidence intervals. Differences in the average differences are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

Accuracy Inter group, which is consistent with this explanation.

We define our third alternative outcome, which we call the Opposite Sharing Rate because its denominator is the complement of the denominator for the primary outcome in Equation 1, as follows:

$$share_i = \frac{\sum_{j=1}^3 1 \left[ S_i^{M,post}(j) | S_i^{T,pre}(j) = 0 \right]}{3 - \sum_{l=j}^3 S_i^{T,pre}(j)} \quad (2)$$

Here the denominator is the number of non-misinformation posts participant  $i$  does NOT intend to share in the pre-survey out of the three they are shown. The numerator is the number of misinformation posts corresponding to the non-misinformation posts participant  $i$  does NOT intend to share in the pre-survey that participant  $i$  intends to share in the post-survey. We drop 4,069 participants who intend to share all non-misinformation posts because then the denominator of this outcome equals zero and is undefined. Because sharing behavior differs between the accuracy nudge groups, we drop more Accuracy After participants than Accuracy Inter participants. Again, this differential dropping is not problematic because accuracy nudge groups are randomized into text message course assignment groups.

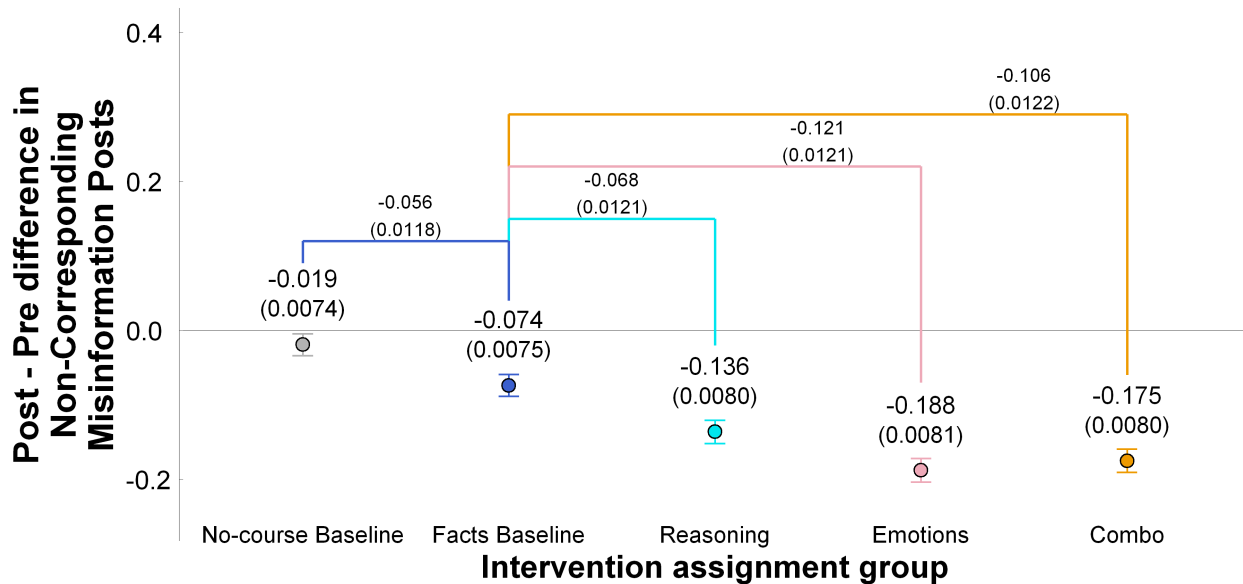


Figure 7: Differences in Non-Corresponding Misinformation Sharing, by Intervention Assignment Group

*Notes:* Sample includes the 8,684 participants who completed the post-survey. Each point displays the average post- minus pre-survey difference in the proportion of misinformation posts shared, excluding those misinformation posts in the post-survey that correspond to non-misinformation posts in the pre-survey, for participants in their respective intervention assignment group, pooling participants in the Accuracy After and Accuracy Inter groups. Above each point, the standard error of each difference is shown in parentheses below the difference. The thin colored bars represent 95% confidence intervals. Differences in the average differences are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

This outcome is motivated by the concern that people may be more likely to share misinformation than non-misinformation, holding constant the relevant fact, *and* that the treatment courses might also differentially affect sharing based on interest in the content versus sharing based on a preference for misinformation over non-misinformation posts. Figure 8 shows that all treatment courses are an improvement over the No-course and Facts baselines using this outcome measure, but the treatment effects are much smaller. This decrease in magnitude is partially due to lower rates in baselines (63.9% on average for the primary outcome versus 29.5% for this outcome), but the treatment effect size in percentage term is also lower. The primary outcome decreases 24% from the Facts baseline to the Emotions course, but the opposite outcome decreases only 17% in the same comparison. This result suggests that it may be more difficult to have an impact with the treatment courses on people who are more likely to share a misinformation version of a post than a non-misinformation version of the same post.

The baseline for this alternative outcome is also a validation of our primary outcome. The implicit assumption in our primary outcome is that, if a participant does not intend to share a fact in a non-misinformation post, they will be less likely to report an intention to share that same fact in a misinformation post as well. If we instead observe that misinformation

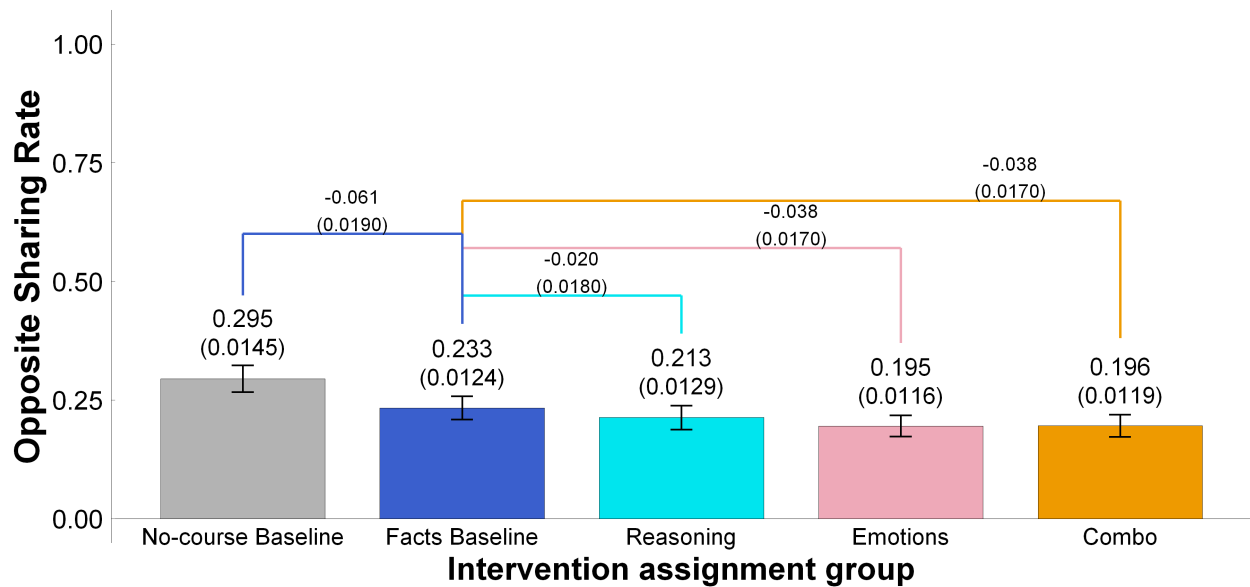


Figure 8: Opposite Outcome, by Intervention Assignment Group

*Notes:* Sample includes the 4,702 participants who completed the post-survey and shared at most two non-misinformation posts in the pre-survey. Each bar displays the Opposite Sharing Rate for misinformation posts, as defined in Equation 2, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Opposite Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Opposite Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

sharing rates are higher on facts that participants do not intend to share in the pre-survey compared to those that participants do intend to share, we would be concerned about the validity of our primary outcome. Comparing Figure 4 to Figure 8 shows that, as expected, participants are more likely to indicate an intention to share misinformation posts about a fact for which they have already indicated an intention to share a non-misinformation post (51.5% on average) compared to facts for which they have already indicated an intention to NOT share a non-misinformation post (22.5% on average), validating the assumption underlying our primary outcome.

All outcomes and tests on sharing behavior support the conclusion that the treatment courses, and the Emotions course in particular, effectively decrease misinformation sharing; however, we might be concerned that these interventions are only stopping people from sharing posts that could be misinformation, but not changing how they think about the misinformation posts. We address this concern with the last alternative outcome using the accuracy scores participants give for the same posts for which they report their sharing decision. Accuracy scores are defined using a numerical encoding of the answer to the “To the best of your knowledge, how accurate is the claim in the above post?” where “Not at all accurate” is encoded as  $-3$  or  $3$  if the post is a non-misinformation post or a misinformation post, respectively; “Not very accurate” is encoded as  $-1$  or  $1$  if the post is a non-misinformation post or a misinformation post, respectively; “Somewhat accurate” is encoded as  $1$  or  $-1$  if the post is a non-misinformation post or a misinformation post, respectively; and “Very accurate” is encoded as  $3$  or  $-3$  if the post is a non-misinformation post or a misinformation post, respectively. Summing this numerical encoding of all posts a participant saw in either the pre- or the post-survey yields accuracy discernment.

Figure 9 shows that the treatment courses increase accuracy discernment over both baselines. The Emotions course again has the largest treatment effect, increasing discernment by  $0.86$  ( $SE = 0.279$ ) compared to  $0.59$  ( $SE = 0.287$ ) in the Combo course and  $0.27$  ( $SE = 0.290$ ) in the Reasoning course. Accuracy discernment in the No-course baseline does not change, but making misinformation salient for five days improves accuracy discernment by  $1.03$  ( $SE = 0.294$ ) in the Facts baseline. In summary, the treatment courses not only decrease misinformation sharing behavior, they decrease how accurate participants perceive misinformation to be.

### 3.3 Long-run effects

We test the long-run effects of the treatment courses in this section. To evaluate the persistent effects of the treatment courses, we again use our primary outcome defined in Equation 1, but instead use the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey. After dropping participants who did not share any non-misinformation post in the follow-up survey, our sample includes 4,251 participants who completed the follow-up survey. Note that again, we have differential dropping between accuracy nudge groups, but these groups are randomized into the treatment courses we evaluate.

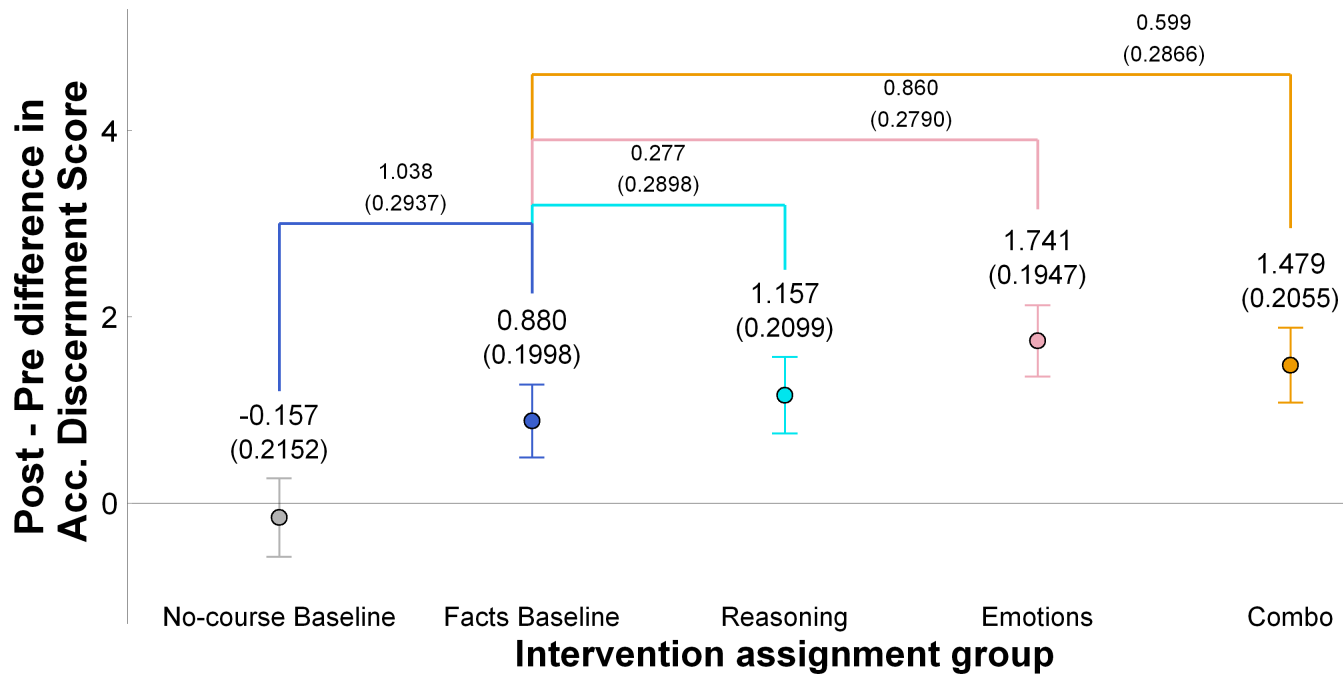


Figure 9: Accuracy Score Discernment, by Intervention Assignment Group

*Notes:* Sample includes the 8,684 participants who completed the post-survey. Each point displays the average post- minus pre-survey difference in accuracy discernment for participants in their respective intervention assignment group, pooling participants in the Accuracy After and Accuracy Inter groups. Above each point, the standard error of each sharing rate is shown in parentheses below the sharing rate. The thin colored bars represent 95% confidence intervals. Differences in the average sharing rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses below.



Figure 10 shows that all treatment courses have a persistent effect on misinformation sharing seven to eleven weeks after course completion. The Emotions course continues to be the most effective, decreasing misinformation sharing by 12.3 p.p. ( $SE = 1.98$  p.p) more than the Facts baseline. In comparison, the Reasoning and Combo courses decrease misinformation sharing by only 6.8 p.p. ( $SE = 1.98$  p.p.) and 8.1 p.p. ( $SE = 2.00$  p.p.) more than the Facts baseline, respectively.

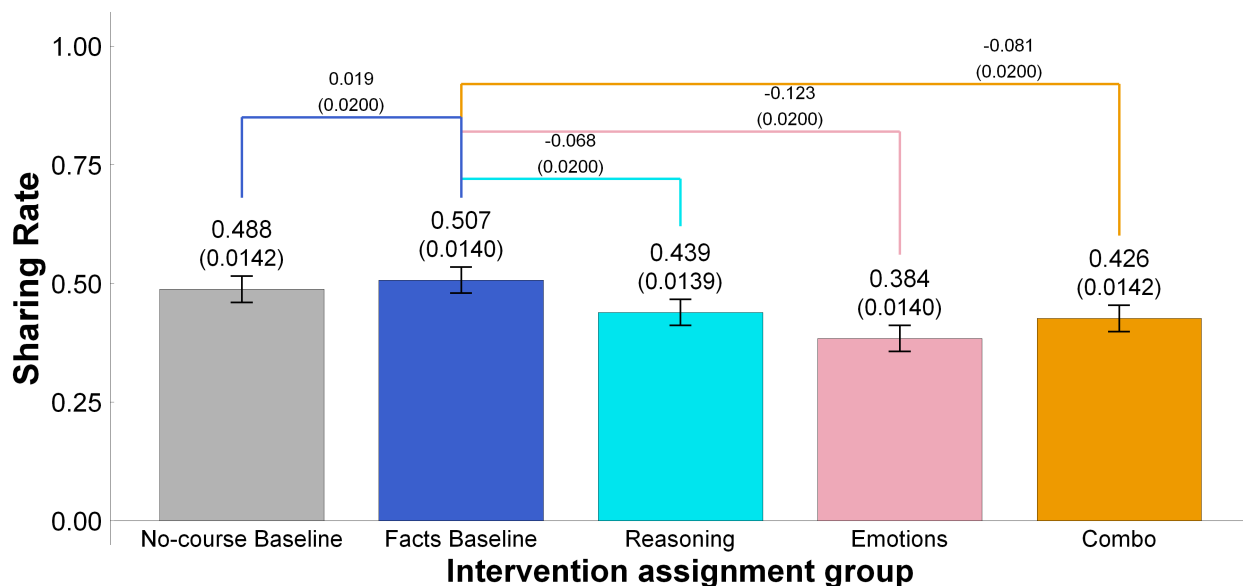


Figure 10: Misinformation Sharing in Follow-up, by Intervention Assignment Group

*Notes:* Sample includes the 4,251 participants who completed the follow-up survey and shared at least one non-misinformation post in the post-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in 1, -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in the average Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

Recall that the No-course baseline group received the Combo course after completing the post-survey, which raises the question of why the Combo course assignment group decreases misinformation sharing 8.18 p.p. ( $SE = 2$  p.p.) more than the No-course baseline even though both are exposed to the same intervention at the time of the Follow-up Experiment. This difference is likely due to low take-up (52%) of the Combo course because payment was already complete. Given that participants in other text message course intervention groups had to complete the text message course prior to completing the post-survey, 100% of the sample in the Combo course assignment group was exposed to the text message course. In comparison, only about half of the sample in the No-course baseline was exposed to the text message course.

We pool across participants assigned to the prime in Figure 10 because the difference in sharing rates between primed and non-primed participants is only 2.0 p.p. ( $SE = 1.82$  p.p.).

The persistent effect of the treatment courses in combination with a null effect of the prime suggests that the treatment courses are effectively and enduringly educating users. Appendix Figures C6 and C7 show the results for the follow-up separately by primed and non-primed participants.

Figure 11 shows that we continue to find that the Emotions course decreases misinformation sharing on Emotions posts more than the other courses (7.95 p.p.,  $SE = 2.90$  p.p., more than the Reasoning course, and 6.93 p.p.,  $SE = 2.90$  p.p., more than the Combo course). The differences between the Emotions course and the other courses on Reasoning and Combo posts have the same estimated direction as in the Main Experiment, except for the difference with the Combo course on the Reasoning posts, as Emotions decreases misinformation sharing by 2.87 ( $SE = 2.96$  p.p.) more than the Combo course in the follow-up survey.

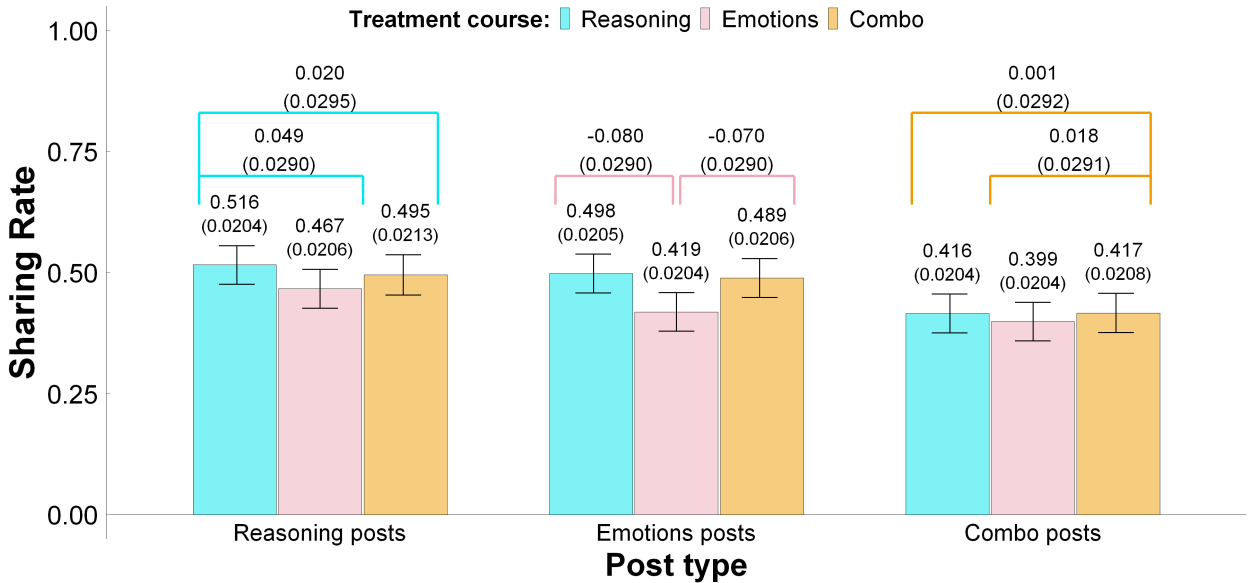


Figure 11: Misinformation Sharing in Follow-up, by Intervention Assignment Group and Post Type

*Notes:* Sample includes the 2,554 participants Reasoning, Emotions or Combo intervention assignment groups who completed the follow-up survey and shared at least one non-misinformation post in the post-survey. Each group of bars displays the Sharing Rate for misinformation posts of each type, as defined in Equation 1 -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After treatments. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in the average Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

### 3.3.1 Reflective Questions

Last, we analyze the responses to two of the reflective questions participants answer in the follow-up survey: “Has the Inoculation against Misinformation course changed your behavior

on social media? If so, how?” (Question 4 or Q4) and “If you were to tell a friend what you learned in the course, what tip would you share?” (Question 5 or Q5). The follow-up survey includes these questions so that we could elicit from participants what they report learning from the course and integrating into their social media behavior. As discussed in the Introduction, arguably the most important limitation of this study is its use of survey sharing outcomes, so we use these free response questions to learn about behavior outside of the survey. This approach has limitations, as we only know what participants are reporting, which could still be subject to experimenter demand.

To start, Table 6 provides summary statistics on all five questions we include in our analysis. Responses are 65.6 characters ( $SE = 0.63$ ), or 12.2 words ( $SE = 0.11$ ), on average for Q4 and 68.1 characters ( $SE = 0.59$ ), or 11.6 words ( $SE = 0.10$ ), for Q5, both of which are longer responses than the other three questions. In general, however, participants wrote a meaningful quantity of text. With the notable exception of Question 3, which asks about participants’ feelings when they see misinformation, participants in the Emotions course wrote the most. On the two questions we include in this analysis, participants in the Emotions course wrote 3.3 characters ( $SE = 1.82$ ) and 2.9 characters ( $SE = 1.91$ ) more compared to the Facts baseline.

	No-course baseline	Facts baseline	Reasoning course	Emotions course	Combo course	Treatment courses	Baselines	All
<b>Number of Characters</b>								
Reflective Question 1	53.107 (1.3825)	54.412 (1.4749)	51.926 (1.3493)	55.739 (1.3358)	54.254 (1.5016)	53.998 (0.8072)	53.792 (1.0152)	53.919 (0.6318)
Reflective Question 2	48.386 (1.4985)	51.960 (1.5060)	49.861 (1.6410)	53.985 (1.5249)	48.586 (1.4022)	50.843 (0.8810)	50.273 (1.0647)	50.625 (0.6795)
Reflective Question 3	45.520 (1.3321)	46.125 (1.3059)	44.477 (1.2470)	46.178 (1.2742)	43.553 (1.2317)	44.750 (0.7227)	45.839 (0.9329)	45.165 (0.5714)
Reflective Question 4	66.011 (1.9498)	64.375 (1.3139)	66.469 (1.2674)	67.475 (1.3368)	65.104 (1.3780)	66.358 (0.7674)	65.148 (1.1530)	65.895 (0.6473)
Reflective Question 5	66.087 (1.3560)	67.639 (1.3264)	69.537 (1.3309)	71.282 (1.4348)	67.421 (1.3405)	69.429 (0.7920)	66.903 (0.9486)	68.465 (0.6092)
<b>Number of Words</b>								
Reflective Question 1	8.304 (0.2322)	8.587 (0.2465)	8.083 (0.2293)	8.769 (0.2227)	8.530 (0.2444)	8.465 (0.1342)	8.453 (0.1700)	8.460 (0.1053)
Reflective Question 2	8.888 (0.2656)	9.525 (0.2652)	9.140 (0.2885)	9.858 (0.2667)	8.850 (0.2463)	9.288 (0.1546)	9.225 (0.1880)	9.264 (0.1195)
Reflective Question 3	8.579 (0.2435)	8.649 (0.2374)	8.372 (0.2250)	8.760 (0.2340)	8.242 (0.2241)	8.461 (0.1316)	8.616 (0.1700)	8.520 (0.1041)
Reflective Question 4	12.231 (0.3228)	12.014 (0.2408)	12.400 (0.2331)	12.613 (0.2443)	12.206 (0.2512)	12.408 (0.1404)	12.117 (0.1985)	12.297 (0.1152)
Reflective Question 5	11.242 (0.2410)	11.582 (0.2358)	11.777 (0.2398)	12.096 (0.2521)	11.470 (0.2387)	11.784 (0.1408)	11.421 (0.1686)	11.645 (0.1083)

Table 6: Summary Statistics for Reflective Questions

*Notes:* Sample includes the 5,316 participants who completed the follow-up survey. Standard errors are in parentheses below means. Questions 1 to 5 are displayed in subsection 2.2.

To learn whether participants in treatment courses are more likely to report behavior changes on social media consistent with treatment, we use keywords like “stop,” “pause,” “question,”

and “evaluate” to define a binary variable based on whether a participant’s response to the question includes one of these keywords. These words correspond to discussing a technique taught in one or more of the courses. Figure 12 shows that these keywords are more likely to show up in responses from participants in one of the treatment courses compared to the Facts baseline.<sup>24</sup> Figure 13 shows that participants in the treatment courses are also more likely to tell their friends a tip with these keywords than participants in the Facts baseline. In summary, participants in the treatment courses are more likely to report behavior consistent with the strategies suggested in the treatment courses than the Facts baseline.

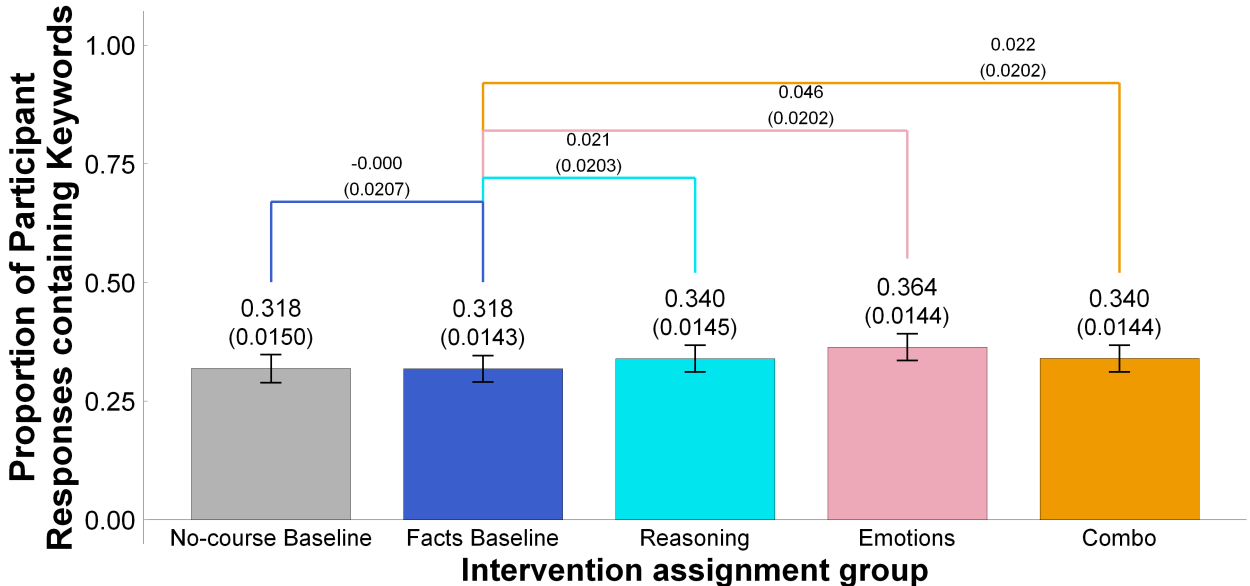


Figure 12: Self-reported Social Media Behavior (Q4) in Follow-up, by Intervention Assignment Group

*Notes:* Sample includes the 5,316 participants who completed the follow-up survey. Each bar displays the proportion of participants whose response to question 4 in the follow-up survey contained one of the keywords. Above each bar, the standard error of each proportion is shown in parentheses below the proportion. The thin black bars represent 95% confidence intervals. Differences in the proportions are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses below.

### 3.4 Accuracy Nudge

In this section, we address three questions. First, does the accuracy nudge effectively decrease misinformation sharing in our study? Second, is the accuracy nudge more or less effective than the treatment courses? Third, are there interaction effects between the accuracy nudge and the treatment courses?

For the analyses in this section, it is helpful to think of our Main Experiment as two separate

<sup>24</sup>It is not surprising that these words also come up in responses from participants in the Facts baseline because the facts about misinformation could induce participants to be more cautious in the same way that the treatment courses actively teach participants to be.

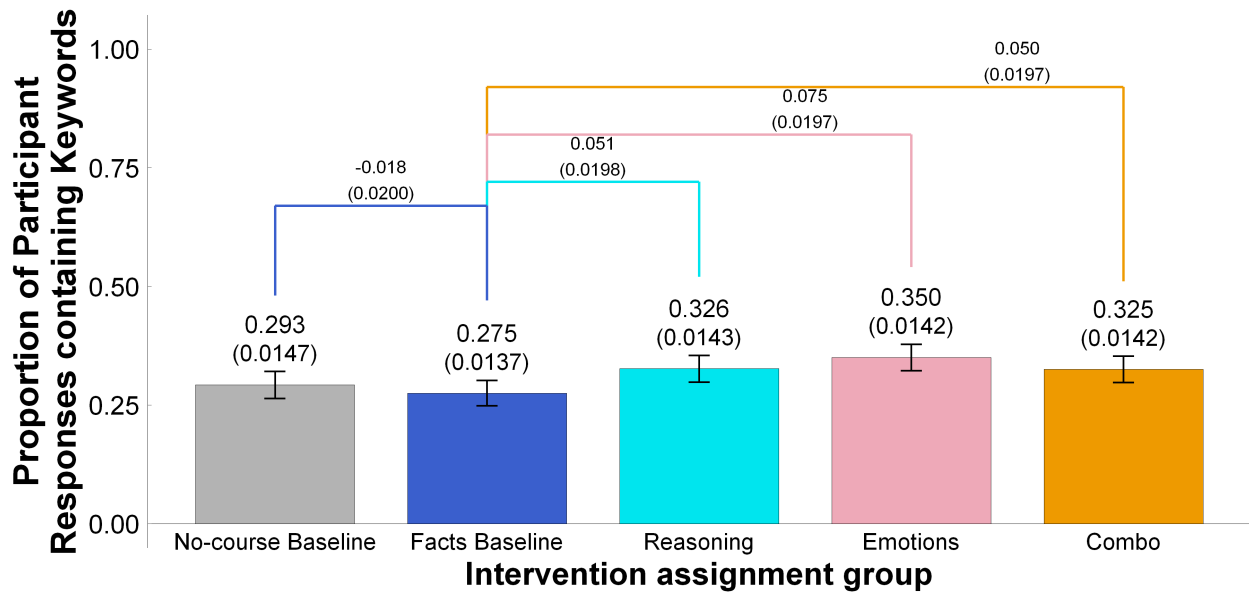


Figure 13: Self-reported Tip to Friend (Q5) in Follow-up, by Intervention Assignment Group

*Notes:* Sample includes the 5,316 participants who completed the follow-up survey. Each bar displays the average proportion of participants whose response to question 5 in the follow-up survey contained one of the keywords. Above each bar, the standard error of each proportion is shown in parentheses below the proportion. The thin black bars represent 95% confidence intervals. Differences in the proportions are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses below.

experiments, one in the pre-survey and one in the post-survey. In the pre-survey experiment, we can compare participants in the Accuracy Inter group who see the accuracy question before each sharing question to participants in the Accuracy After group who have not see any accuracy questions when they answer the sharing questions (see Figure 2). From this experiment, we can analyze the treatment effect of asking participants about the accuracy of a post before asking them for a sharing decision compared to not asking participants about accuracy at all.

In the second experiment in the post-survey, all participants have been exposed to accuracy questions in the pre-survey. In addition, some have seen text message courses. By analyzing outcomes at this point in the experiment, we can learn the treatment effect of the treatment courses combined with different levels of exposure to accuracy questions. Participants in the Accuracy After group again do not see the accuracy questions until after they answer the sharing questions, but because they are exposed to accuracy questions in the pre-survey, we learn from this group the treatment effect of the treatment courses combined with having seen ten accuracy questions five days earlier.<sup>25</sup> Participants in the Accuracy Inter group have prior exposure to interweaved accuracy questions in the pre-survey, and again see the accuracy questions interweaved during the post-survey.

<sup>25</sup>To our knowledge, there is no evidence on the durability of treatment effects from accuracy nudges.

We focus on an alternative outcome in this section to evaluate the accuracy nudge. The outcome in Equation 1 is defined based on a change between pre- and post-survey behavior, but participants assigned to see the accuracy nudge see the Accuracy Inter treatment in both surveys. Thus, a change in behavior between the pre- and post-survey is not a relevant metric for evaluating the nudge. We instead consider as our primary outcome the proportion of misinformation posts (out of six) a participant intends to share in the pre-survey  $misinfo_i^{pre}$  or in the post-survey  $misinfo_i^{post}$ .

Table 5 shows that the accuracy nudge effectively decreases misinformation sharing. In the pre-survey, averaging over all text courses, participants asked to rate the accuracy of a post prior to making a sharing decision share 51.2% ( $SE = 0.50$ ), or three of the six misinformation posts. Participants in the Accuracy After group who do not see the accuracy questions before making their sharing decision share 6.7 p.p. ( $SE = 0.71$  p.p.) more misinformation, or about half a post. So, the high-dosage version of the accuracy nudge we implement changes misinformation sharing behavior in our context.

We can also evaluate the accuracy nudge using the post-survey. At this point in the experiment, participants have potentially been exposed to text message courses, and also have had previous exposure to accuracy questions (either interweaved in Accuracy Inter or at the end in Accuracy After). For simplicity, consider the No-Course baseline group. Then, the difference in post-survey sharing outcomes between Accuracy Inter and Accuracy After groups compares a group that saw interweaved accuracy five days earlier in the pre-survey, and then again in the post-survey, to a group that saw ten accuracy questions five days earlier in the pre-survey, but had not see accuracy questions in the post-survey at the point in which they answered the sharing questions. Table 5 shows that the Accuracy Inter group in the No-course baseline shared 51.4% ( $SE = 1.22$ ) of misinformation posts in the post-survey relative to 57.8% ( $SE = 1.11$ ) in the Accuracy After group, yielding a treatment effect of 6.4 p.p. ( $SE = 1.65$ ) for the accuracy nudge, which is similar but slightly smaller than the pre-survey treatment effect. Appendix B shows that qualitatively similar results hold for pre-survey and post-survey sharing and accuracy discernment.

We next compare the estimated treatment effect of the accuracy nudge to the estimated treatment effect of the treatment courses to learn whether the treatment courses are more or less effective than the accuracy nudge. Because the 18 p.p. treatment effect of the courses we estimate in subsection 3.1 uses a different outcome (Equation 1) than we use to estimate the treatment effect of the accuracy nudge in this subsection, we need to estimate an alternative treatment effect for the treatment courses using the same outcome measure: post-survey misinformation sharing.

To estimate the alternative treatment effect of the treatment courses, we restrict our sample to the 3,454 participants in the Accuracy After group and in one of the treatment courses (Emotions, Reasoning, or Combo) or the No-course baseline. Participants in this group are all exposed to ten accuracy questions in the pre-survey. Taking the difference between the post-survey misinformation sharing of participants exposed to one of the treatment courses and participants exposed to no course yields the treatment effect of the courses (in the absence of Accuracy Inter, but for participants who have previously been exposed to ten Accuracy questions in the pre-survey).

Table 5 shows that participants in the No-course baseline, Accuracy After assignment group share 57.8% ( $SE = 1.11$ ) of the six misinformation posts in the post-survey. Participants in the Accuracy After group and one of the treatment course assignment groups share 18.0 p.p. ( $SE = 1.28$  p.p.) less misinformation (31.1% of baseline outcome); and this treatment effect estimate is very similar to the 18.1 p.p. (28% of baseline outcome) estimated using the Sharing Rate, our primary outcome for treatment course evaluation (baselines are also similar).

Since this treatment effect of the treatment courses is estimated for participants who were exposed to ten accuracy questions in the pre-survey (five days earlier), it is most meaningful to compare it to the treatment effect of the accuracy nudge for participants with similar prior experience. Thus, we compare the treatment course treatment effect to the treatment effect of the accuracy nudge on post-survey outcomes measured for those individuals in the No-course baseline. As described above, this effect is 6.4 p.p., much smaller than the 18.0 p.p. treatment effect for the text message courses, recalling that the control group for both treatment effects is the same at 57.8%.

Last, we test whether there is an interaction effect between the accuracy nudge and the treatment courses; that is, whether the two treatments are substitutes, complements, or have independent effects. To estimate this incremental effect, we take the difference of two differences, the first of which we already estimated above. The first difference estimates the difference between the two accuracy nudge groups in the post-survey when they are not exposed to a treatment course:  $misinfo_i^{post}$  in the Accuracy Inter, No-course baseline group minus  $misinfo_i^{post}$  in the Accuracy After, No-course baseline group. The second difference estimates the difference between the two accuracy nudge groups in the post-survey when they are exposed to one of the treatment courses:  $misinfo_i^{post}$  in the Accuracy Inter, treatment course group minus  $misinfo_i^{post}$  in the Accuracy After, treatment course group. The difference in these two differences captures whether asking accuracy interweaved in the sharing questions after educating participants about techniques to combat misinformation decreases misinformation more or less than the accuracy nudge alone.

We find a smaller difference between the accuracy nudge groups exposed to the treatment courses compared to the difference between the accuracy nudge groups not exposed to the treatment courses, suggesting that the accuracy nudge is not more valuable in combination with the treatment courses than alone, as illustrated in Figure 14. As estimated above, the difference in  $misinfo_i^{post}$  between the two accuracy nudge groups in the No-course baseline is 6.4 p.p. ( $SE = 1.65$  p.p.). The same difference in  $misinfo_i^{post}$  between the two accuracy nudge groups exposed to the treatment courses is 4.5 p.p. ( $SE = 0.90$  p.p.). The difference in differences is 1.9 p.p. ( $SE = 1.88$  p.p.), so we can bound the incremental effect of the nudge and the courses at a 1.8 p.p. increase in the sharing rate, with 95% confidence.

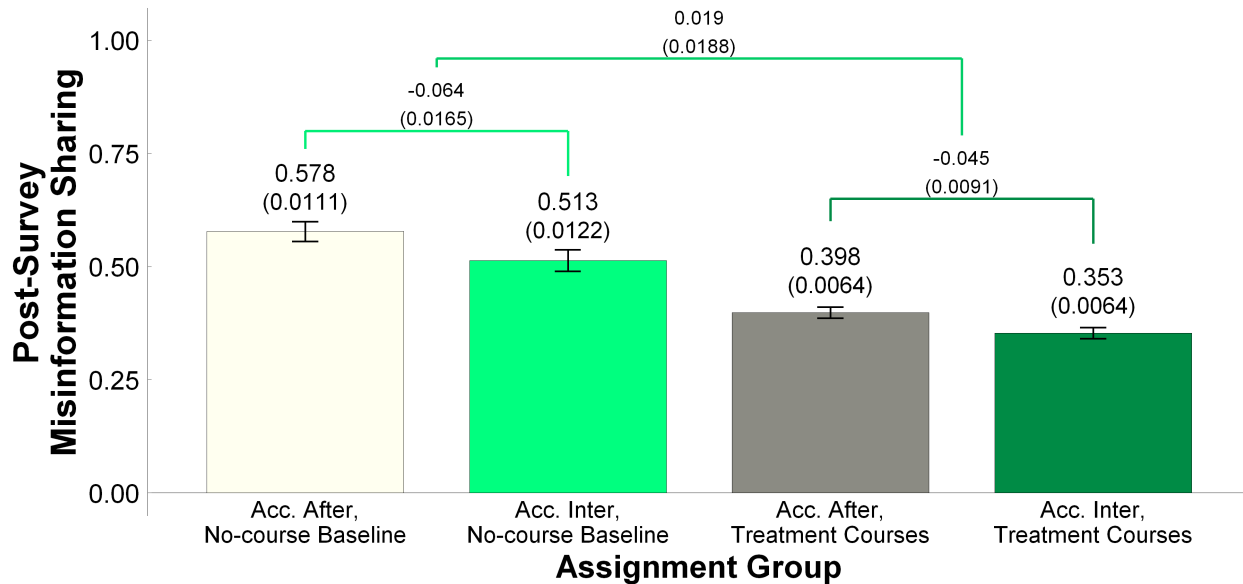


Figure 14: Effect of Accuracy Nudge and Treatment Courses

*Notes:* Sample includes the 6,894 participants in the Reasoning, Emotions, Combo, and No-course baseline assignment groups who completed the post-survey. Each bar displays the proportion of misinformation posts shared in the post-survey, by participants in their respective assignment group. Above each bar, the standard error of each sharing rate is shown in parentheses below the sharing rate. The thin black bars represent 95% confidence intervals. The differences in misinformation sharing proportions, along with their difference, are shown above lines connecting the two relevant assignment groups, with standard errors in parentheses below.

## 4 Discussion

This paper evaluates three versions of a technique-based text message course on misinformation in a field experiment in Kenya to understand whether treating reasoning- or emotion-based techniques is most effective and whether the two approaches combined are complementary. Overall, we find that the expert-developed text message courses effectively decrease misinformation sharing both immediately after completing the course and seven to eleven weeks later. We also show that a high-dosage accuracy nudge decreases misinformation sharing, but not as much as the treatment courses. We find no evidence of complementarities from adding the accuracy nudge to the treatment courses, further confirming the preeminence of teaching about emotion-based techniques. Last, we show that participants in the treatment courses are more likely to report using (and telling friends about) strategies taught in the course than those not exposed to the treatment courses. The long-run effects of the course, its performance relative to the accuracy nudge, and the self-reported behavior in the follow-up survey all support the generalizability of our conclusions. At the same time, future research that tests these interventions using on-platform behavior, as opposed to the survey behavior we and the vast majority of the literature use, is an important next step.



Our finding that the Emotions course is consistently the most effective course suggests that people make impulsive, emotions-driven sharing decisions that lead to misinformation sharing. This finding has three concerning implications. First, heightened emotional states are correlated with greater belief in misinformation (Martel et al., 2020). Second, the misinformation literature to date focuses primarily on evaluating interventions that counter reasoning-based techniques, like debunking interventions and technique-based interventions to determine whether a post is misinformation, rather than identifying or developing interventions that counter emotion-based techniques.<sup>26</sup> Last, posts that contain emotional language are dispersed more broadly and quickly than posts without emotional language based on text analyses of social media content (Brady et al., 2017; Pröllochs et al., 2021). The importance of this last point is further underscored by our results showing that posts with emotional language are precisely the types of posts for which the Emotions course has the greatest advantage over the other courses. So, not only is the Emotions course the most effective overall, it has the greatest advantage in changing behavior on the types of posts that are most concerning for promoting the belief in and the spread of misinformation.

One way to understand the psychological mechanisms driving the success of the Emotions course is through the influential Elaboration Likelihood Model (ELM) of persuasion (Petty and Cacioppo, 1986) in the psychology literature. The ELM posits that people use either central route processing, a high level of thinking (elaboration) focused on the merits of a persuasive argument, or peripheral route processing, a low level of thinking that uses environmental cues and heuristics, to evaluate a persuasive argument (or some combination) (Petty and Hinskamp, 2017). One finding in this body of literature based on lab studies is that emotions can affect how much central versus peripheral elaboration a person does when evaluating a persuasive argument and, when peripheral route processing dominates, whether the person ultimately accepts or rejects the argument (see Petty and Briñol (2015) for a review).

The primary technique taught in the Emotions course is to “Stop and Question” content that elicits a strong emotional reaction, which is a type of emotion regulation through “cognitive reappraisal.” The emotions regulation literature in psychology finds in lab experiments and observational studies that cognitive reappraisal is frequently successful at changing perceived emotions in the short-run through changing how a person thinks about the emotions they are experiencing. Cognitive reappraisal also is more successful in the long-run than other emotion regulations strategies like suppression (see McRae and Gross (2020) for a review). Thus, one way to understand how the Emotions course might work psychologically is it teaches participants that misinformation often uses emotions to manipulate them (e.g., by triggering peripheral instead of central route processing). Then, the Emotions course provides a strategy participants can use to “reappraise” how they think about those emotions when they experience them. This reappraisal could trigger a central route process that leads to higher elaboration and could also lead participants to evaluate a post based on the merits of its argument rather than heuristics.

---

<sup>26</sup>One exception is Bago et al. (2022), which tests whether emotion regulation techniques, such as suppression (by prompting participants to not let their own feeling show) and cognitive reappraisal (by prompting participants to adopt a detached and unemotional attitude), would reduce reported belief in misinformation; however, they find little evidence.

One important direction for future work is to evaluate the effectiveness of this and other interventions to combat the misinformation sharing using on-platform behavior. Another direction for future work is to understand the precise mechanisms driving emotional sharing on social media. We propose that emotion-based techniques operates through the first-stage of a two-stage decision process in which social media users must first stop to evaluate whether a post is misinformation before they can conduct the evaluation, but several mechanisms could be at work here. For example, teaching users about emotion-based techniques may lead them to stop to evaluate a post that contains emotional language either because 1) the user intellectually identifies the emotional language and uses that as a signal for a stopping mechanism, and/or 2) the user feels strong emotions and uses that as a signal for a stopping mechanism. Our study provides some evidence for the second explanation as the Emotions course was as effective as the Reasoning or Combo course for Reasoning posts, but a research design to identify mechanisms would be better suited to answer this question.

## References

- Arechar, A. A., J. N. L. Allen, a. berinsky, R. Cole, Z. Epstein, K. Garimella, A. Gully, J. G. Lu, R. M. Ross, M. Stagnaro, and et al. (2022, Feb). Understanding and reducing online misinformation across 16 countries on six continents.
- Armanasco, A. A., Y. D. Miller, B. S. Fjeldsoe, and A. L. Marshall (2017). Preventive health behavior change text message interventions: a meta-analysis. *American journal of preventive medicine* 52(3), 391–402.
- Athey, S., M. Cersosimo, K. Koutout, and Z. Li (2022, July). Inoculation against misinformation.
- Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics* 47(2), 1148–1178.
- Bago, B., L. R. Rosenzweig, A. J. Berinsky, and D. G. Rand (2022). Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. *Cognition and Emotion* 36(6), 1166–1180.
- Basol, M., J. Roozenbeek, M. Berriche, F. Uenal, W. P. McClanahan, and S. v. d. Linden (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against covid-19 misinformation. *Big Data & Society* 8(1), 20539517211013868.
- Basol, M., J. Roozenbeek, and S. Van der Linden (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition* 3(1).
- Berlinski, N., M. Doyle, A. M. Guess, G. Levy, B. Lyons, J. M. Montgomery, B. Nyhan, and J. Reifler (2021). The effects of unsubstantiated claims of voter fraud on confidence in elections. *Journal of Experimental Political Science*, 1–16.

- Brady, W. J., J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114(28), 7313–7318.
- Bryan, C. J., D. S. Yeager, C. P. Hinojosa, A. Chabot, H. Bergen, M. Kawamura, and F. Steubing (2016). Harnessing adolescent values to motivate healthier eating. *Proceedings of the National Academy of Sciences* 113(39), 10830–10835.
- Castleman, B. L. and L. C. Page (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates? *Journal of Economic Behavior & Organization* 115, 144–160.
- Compton, J., S. van der Linden, J. Cook, and M. Basol (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass* 15(6), e12602.
- Cook, J. (2020). Deconstructing climate science denial. *Research handbook on communicating climate change*, 62–78.
- Cook, J., S. Lewandowsky, and U. K. Ecker (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLoS one* 12(5), e0175799.
- Gavin, L., J. McChesney, A. Tong, J. Sherlock, L. Foster, and S. Tomsa (2022). Fighting the spread of covid-19 misinformation in kyrgyzstan, india, and the united states: How replicable are accuracy nudge interventions? *Technology, Mind, and Behavior*, No–Pagination.
- Guess, A. M., M. Lerner, B. Lyons, J. M. Montgomery, B. Nyhan, J. Reifler, and N. 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), 15536–15545.
- Harjani, T., M.-S. Basol, J. Roozenbeek, and S. van der Linden (2023). Gamified inoculation against misinformation in india: A randomized control trial. *Journal of Trial & Error*.
- Ho, K. K., J. Y. Chan, and D. K. Chiu (2022). Fake news and misinformation during the pandemic: what we know and what we do not know. *IT Professional* 24(2), 19–24.
- Lewandowsky, S. and S. Van Der Linden (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology* 32(2), 348–384.
- List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.
- Maertens, R., J. Roozenbeek, M. Basol, and S. van der Linden (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied* 27(1), 1.
- Martel, C., G. Pennycook, and D. G. Rand (2020). Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications* 5(1), 1–20.

- McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology* 63(2), 326.
- McRae, K. and J. J. Gross (2020). Emotion regulation. *Emotion* 20(1), 1.
- Nguyen, T. and S. Cecchini (2021). Countering covid-19 misinformation in africa. <https://www.thinkglobalhealth.org/article/countering-covid-19-misinformation-africa>.
- Offer-Westort, M., L. R. Rosenzweig, and S. Athey (2021). Optimal policies to battle the coronavirus “infodemic” among social media users in sub-saharan africa. *OSF Registered Study*.
- Pennycook, G., Z. Epstein, M. Mosleh, A. A. Arechar, D. Eckles, and D. G. Rand (2021). Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855), 590–595.
- Petty, R. and L. Hinsenkamp (2017). The sage encyclopedia of political behavior.
- Petty, R. E. and P. Briñol (2015). Emotion and persuasion: Cognitive and meta-cognitive processes impact attitudes. *Cognition and Emotion* 29(1), 1–26.
- Petty, R. E. and J. T. Cacioppo (1986). The elaboration likelihood model of persuasion. In *Communication and persuasion*, pp. 1–24. Springer.
- Pröllochs, N., D. Bär, and S. Feuerriegel (2021). Emotions explain differences in the diffusion of true vs. false social media rumors. *Scientific reports* 11(1), 1–12.
- Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* 89(427), 846–866.
- Rodgers, A., T. Corbett, D. Bramley, T. Riddell, M. Wills, R.-B. Lin, and M. Jones (2005). Do u smoke after txt? results of a randomised trial of smoking cessation using mobile phone text messaging. *Tobacco control* 14(4), 255–261.
- Romano, J. P. and M. Wolf (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics* 35(4), 1378–1408.
- Roozenbeek, J., R. Maertens, W. McClanahan, and S. van der Linden (2021). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement* 81(2), 340–362.
- Roozenbeek, J. and S. Van Der Linden (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of risk research* 22(5), 570–580.
- Roozenbeek, J., S. Van Der Linden, B. Goldberg, S. Rathje, and S. Lewandowsky (2022). Psychological inoculation improves resilience against misinformation on social media. *Science advances* 8(34), eabo6254.
- Suffoletto, B., J. Kristan, C. Callaway, K. H. Kim, T. Chung, P. M. Monti, and D. B. Clark (2014). A text message alcohol intervention for young adult emergency department patients: a randomized clinical trial. *Annals of emergency medicine* 64(6), 664–672.

- Telzer, E. H., A. J. Fuligni, M. D. Lieberman, and A. Galván (2014). Neural sensitivity to eudaimonic and hedonic rewards differentially predict adolescent depressive symptoms over time. *Proceedings of the National Academy of Sciences* 111(18), 6600–6605.
- van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28(3), 460–467.
- Van der Linden, S., A. Leiserowitz, S. Rosenthal, and E. Maibach (2017). Inoculating the public against misinformation about climate change. *Global Challenges* 1(2), 1600008.
- Yadlowsky, S., S. Fleming, N. Shah, E. Brunskill, and S. Wager (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*.
- York, B. N., S. Loeb, and C. Doss (2019). One step at a time the effects of an early literacy text-messaging program for parents of preschoolers. *Journal of Human Resources* 54(3), 537–566.

## A Tests with Multiple Hypothesis Test Correction

In this appendix, we report the execution of tests from our pre-analysis plan and tests on our non-prespecified Follow-up Experiment. Since we execute multiple tests to evaluate the same hypothesis (e.g., “Do the treatment courses decrease misinformation sharing?” and “Which treatment course is most effective at decreasing misinformation sharing?”), we want to control the probability of drawing at least one false conclusion.

To control the Family-Wise Error Rate (FWER), we correct for multiple hypothesis testing by using the Romano-Wolf procedure (Romano and Wolf, 2007) to adjust p-values. These adjusted p-values are an extension of resample-based p-values for single hypothesis testing, which are defined as the fraction of resamples that produce a Studentized null statistic with a value that is more extreme than the test statistic from the test with the original sample. The Romano-Wolf p-values are defined as the fraction of resamples that produce a “max”-statistic that is more extreme than the test statistic from the test with the original sample, where the “max”-statistic is defined, for a given test  $T$ , as the maximum of the Studentized null statistics from the set of Studentized null statistics corresponding to  $T$  and the tests for which the statistic with the original data are less extreme than the statistic from  $T$ . By capturing the correlation between test statistics through the use of the “max”-statistics, this procedure achieves greater power than other corrections such as Bonferroni or Holm.

We group our hypotheses into five families for this correction. The first family of hypotheses tests whether the treatment courses (Emotions, Reasoning, and Combo) are more effective than the baselines (No-course and Facts). The second family of hypotheses tests which of the treatment courses works best and which work best for each type of misinformation post. The third family of hypotheses tests the effects of the accuracy nudge, separately, and then relative to and in combination with the treatment courses. The fourth family of hypotheses tests whether the treatment courses are more effective than the Facts baseline in the follow-up survey to test the long-run effects of the treatment courses. Last, the fifth family of hypotheses tests which of the treatment courses work best in the long-run, and which work best for each type of misinformation post in the follow-up survey.

## A.1 Family 1 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>No-course</b>	4,648	1,456	0.459 (0.0059)	0.639 (0.0100)	-0.180 (0.0116)	0.0010 [0.0000]
<b>Treatment courses</b>	<b>Facts</b>	4,648	1,584	0.459 (0.0059)	0.563 (0.0103)	-0.104 (0.0118)	0.0001 [0.0000]
<b>Reasoning</b>	<b>No-course</b>	1,502	1,456	0.496 (0.0103)	0.639 (0.0100)	-0.143 (0.0144)	0.0001 [0.0000]
<b>Emotions</b>	<b>No-course</b>	1,617	1,456	0.430 (0.0098)	0.639 (0.0100)	-0.209 (0.0140)	0.0001 [0.0000]
<b>Combo</b>	<b>No-course</b>	1,529	1,456	0.454 (0.0105)	0.639 (0.0100)	-0.185 (0.0145)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	1,502	1,584	0.496 (0.0103)	0.563 (0.0103)	-0.067 (0.0146)	0.0001 [0.0000]
<b>Emotions</b>	<b>Facts</b>	1,617	1,584	0.430 (0.0098)	0.563 (0.0103)	-0.133 (0.0142)	0.0001 [0.0000]
<b>Combo</b>	<b>Facts</b>	1,529	1,584	0.454 (0.0105)	0.563 (0.0103)	-0.109 (0.0147)	0.0001 [0.0000]
<b>Facts</b>	<b>No-course</b>	1,584	1,456	0.563 (0.0103)	0.639 (0.0100)	-0.076 (0.0144)	0.0001 [0.0000]

Table A1: Average Treatment Effects: Misinformation Sharing

*Notes:* Sample includes the 7,688 participants who completed the post-survey and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Sharing Rate for misinformation posts, as defined in 1. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 1 tests using 10000 simulations and therefore have a minimum of 0.0001.

Posts Shared	Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
1	<b>Treatment courses</b>	<b>No-course</b>	786	238	0.336 (0.0169)	0.454 (0.0323)	-0.118 (0.0365)	0.0061 [0.0007]
2	<b>Treatment courses</b>	<b>No-course</b>	1,427	453	0.398 (0.0101)	0.586 (0.0177)	-0.188 (0.0204)	0.0001 [0.0000]
3	<b>Treatment courses</b>	<b>No-course</b>	2,435	765	0.535 (0.0076)	0.729 (0.0113)	-0.193 (0.0136)	0.0001 [0.0000]
1	<b>Treatment courses</b>	<b>Facts</b>	786	283	0.336 (0.0169)	0.406 (0.0292)	-0.070 (0.0338)	0.0628 [0.0187]
2	<b>Treatment courses</b>	<b>Facts</b>	1,427	519	0.398 (0.0101)	0.487 (0.0179)	-0.089 (0.0205)	0.0002 [0.0000]
3	<b>Treatment courses</b>	<b>Facts</b>	2,435	782	0.535 (0.0076)	0.671 (0.0123)	-0.135 (0.0144)	0.0001 [0.0000]
1	<b>Reasoning</b>	<b>No-course</b>	236	238	0.403 (0.0320)	0.454 (0.0323)	-0.051 (0.0455)	0.2967 [0.1303]
2	<b>Reasoning</b>	<b>No-course</b>	466	453	0.430 (0.0177)	0.586 (0.0177)	-0.156 (0.0250)	0.0001 [0.0000]
3	<b>Reasoning</b>	<b>No-course</b>	800	765	0.562 (0.0130)	0.729 (0.0113)	-0.166 (0.0172)	0.0001 [0.0000]
1	<b>Emotions</b>	<b>No-course</b>	280	238	0.300 (0.0274)	0.454 (0.0323)	-0.154 (0.0424)	0.0013 [0.0002]
2	<b>Emotions</b>	<b>No-course</b>	497	453	0.385 (0.0166)	0.586 (0.0177)	-0.201 (0.0243)	0.0001 [0.0000]
3	<b>Emotions</b>	<b>No-course</b>	840	765	0.500 (0.0128)	0.729 (0.0113)	-0.229 (0.0171)	0.0001 [0.0000]
1	<b>Combo</b>	<b>No-course</b>	270	238	0.315 (0.0283)	0.454 (0.0323)	-0.139 (0.0430)	0.0061 [0.0007]
2	<b>Combo</b>	<b>No-course</b>	464	453	0.379 (0.0180)	0.586 (0.0177)	-0.207 (0.0252)	0.0001 [0.0001]
3	<b>Combo</b>	<b>No-course</b>	795	765	0.545 (0.0134)	0.729 (0.0113)	-0.184 (0.0176)	0.0001 [0.0000]
1	<b>Reasoning</b>	<b>Facts</b>	236	283	0.403 (0.0320)	0.406 (0.0292)	-0.004 (0.0433)	0.4729 [0.4649]
2	<b>Reasoning</b>	<b>Facts</b>	466	519	0.430 (0.0177)	0.487 (0.0179)	-0.057 (0.0251)	0.0613 [0.0116]
3	<b>Reasoning</b>	<b>Facts</b>	800	782	0.562 (0.0130)	0.671 (0.0123)	-0.108 (0.0179)	0.0001 [0.0000]
1	<b>Emotions</b>	<b>Facts</b>	280	283	0.300 (0.0274)	0.406 (0.0292)	-0.106 (0.0401)	0.0274 [0.0041]
2	<b>Emotions</b>	<b>Facts</b>	497	519	0.385 (0.0166)	0.487 (0.0179)	-0.102 (0.0244)	0.0002 [0.0000]
3	<b>Emotions</b>	<b>Facts</b>	840	782	0.500 (0.0128)	0.671 (0.0123)	-0.171 (0.0177)	0.0001 [0.0000]
1	<b>Combo</b>	<b>Facts</b>	270	283	0.315 (0.0283)	0.406 (0.0292)	-0.092 (0.0407)	0.0613 [0.0125]
2	<b>Combo</b>	<b>Facts</b>	464	519	0.379 (0.0180)	0.487 (0.0179)	-0.108 (0.0253)	0.0002 [0.0000]
3	<b>Combo</b>	<b>Facts</b>	795	782	0.545 (0.0134)	0.671 (0.0123)	-0.126 (0.0182)	0.0001 [0.0000]
1	<b>Facts</b>	<b>No-course</b>	283	238	0.406 (0.0292)	0.454 (0.0323)	-0.047 (0.0436)	0.2967 [0.1387]
2	<b>Facts</b>	<b>No-course</b>	519	453	0.487 (0.0179)	0.586 (0.0177)	-0.099 (0.0251)	0.0004 [0.0000]
3	<b>Facts</b>	<b>No-course</b>	782	765	0.671 (0.0123)	0.729 (0.0113)	-0.058 (0.0167)	0.0025 [0.0003]

Table A2: Average Treatment Effects: Misinformation Sharing by Number of Non-Misinformation Posts Shared in the Pre-survey



*Notes:* Sample includes the 7,688 participants who completed the post-survey and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy After and Accuracy Inter treatments. As indicated in the first column of the table, the participants are divided into sharing groups according to the number of non-misinformation posts shared in the pre-survey. Number of participants per group: 1 post shared: 1,307; 2 posts shared: 2,399; 3 posts shared: 3,982. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Sharing Rate for misinformation posts, as defined in 1. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 1 tests using 10000 simulations and therefore have a minimum of 0.0001.

## A.2 Family 2 Tests

Post type	Group 1	Group 2	N Obs 1	N Obs 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
All	<b>Reasoning</b>	<b>Emotions</b>	1,502	1,617	0.496 (0.0103)	0.430 (0.0098)	0.066 (0.0143)	0.0001 [0.0000]
All	<b>Reasoning</b>	<b>Combo</b>	1,502	1,529	0.496 (0.0103)	0.454 (0.0105)	0.042 (0.0147)	0.0282 [0.0039]
All	<b>Emotions</b>	<b>Combo</b>	1,617	1,529	0.430 (0.0098)	0.454 (0.0105)	-0.024 (0.0144)	0.3547 [0.0977]
Reasoning	<b>Reasoning</b>	<b>Emotions</b>	1,181	1,261	0.539 (0.0145)	0.509 (0.0141)	0.030 (0.0202)	0.3685 [0.1347]
Reasoning	<b>Reasoning</b>	<b>Combo</b>	1,181	1,202	0.539 (0.0145)	0.489 (0.0144)	0.050 (0.0205)	0.0691 [0.0142]
Reasoning	<b>Emotions</b>	<b>Combo</b>	1,261	1,202	0.509 (0.0141)	0.489 (0.0144)	0.020 (0.0202)	0.4693 [0.3228]
Emotions	<b>Emotions</b>	<b>Reasoning</b>	1,254	1,192	0.454 (0.0141)	0.533 (0.0145)	-0.079 (0.0202)	0.0010 [0.0000]
Emotions	<b>Emotions</b>	<b>Combo</b>	1,254	1,195	0.454 (0.0141)	0.510 (0.0145)	-0.057 (0.0202)	0.0315 [0.0025]
Emotions	<b>Reasoning</b>	<b>Combo</b>	1,192	1,195	0.533 (0.0145)	0.510 (0.0145)	0.022 (0.0205)	0.4693 [0.2766]
Combo	<b>Combo</b>	<b>Reasoning</b>	1,195	1,192	0.454 (0.0145)	0.480 (0.0145)	-0.027 (0.0205)	0.3949 [0.0958]
Combo	<b>Combo</b>	<b>Emotions</b>	1,195	1,254	0.454 (0.0145)	0.403 (0.0137)	0.050 (0.0199)	0.0691 [0.0119]
Combo	<b>Reasoning</b>	<b>Emotions</b>	1,195	1,279	0.480 (0.0145)	0.403 (0.0137)	0.077 (0.0199)	0.0013 [0.0001]

Table A3: Average Treatment Effects: Misinformation Sharing by type of post

*Notes:* Sample includes the 4,648 participants in the Reasoning, Emotions or Combo intervention assignment groups who completed the post-survey and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy After and Accuracy Inter treatments. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Sharing Rate for misinformation posts, as defined in 1, subset to the post type indicated in the first column of the table. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 2 tests using 10000 simulations and therefore have a minimum of 0.0001.

### A.3 Family 3 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>Facts</b>	2,554	857	0.416 (0.0081)	0.507 (0.0140)	-0.091 (0.0162)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	857	857	0.439 (0.0139)	0.507 (0.0140)	-0.068 (0.0198)	0.0005 [0.0003]
<b>Emotions</b>	<b>Facts</b>	865	857	0.384 (0.0140)	0.507 (0.0140)	-0.123 (0.0198)	0.0001 [0.0000]
<b>Combo</b>	<b>Facts</b>	832	857	0.426 (0.0142)	0.507 (0.0140)	-0.081 (0.0200)	0.0003 [0.0000]
<b>Facts</b>	<b>No-course</b>	857	840	0.507 (0.0140)	0.488 (0.0142)	0.019 (0.0200)	0.3414 [0.3392]

Table A4: Average Treatment Effects - Follow-up

*Notes:* Sample includes the 4,251 participants who completed the follow-up survey and shared at least one non-misinformation post in the post-survey, pooling participants in the Accuracy After and Accuracy Inter treatments. Each row displays the results of t-tests. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Sharing Rate for misinformation posts, as defined in 1, -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 3 tests using 10000 simulations and therefore have a minimum of 0.0001.

## A.4 Family 4 Tests

Post type	Group 1	Group 2	N Obs 1	N Obs 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
All	<b>Reasoning</b>	<b>Emotions</b>	1,502	1,617	0.439 (0.0139)	0.384 (0.0140)	0.055 (0.0197)	0.0548 [0.0054]
All	<b>Reasoning</b>	<b>Combo</b>	1,502	1,529	0.439 (0.0139)	0.426 (0.0142)	0.013 (0.0199)	0.9650 [0.5118]
All	<b>Emotions</b>	<b>Combo</b>	1,617	1,529	0.384 (0.0140)	0.426 (0.0142)	-0.042 (0.0199)	0.2396 [0.0355]
Reasoning	<b>Reasoning</b>	<b>Emotions</b>	599	587	0.516 (0.0204)	0.467 (0.0206)	0.049 (0.0290)	0.4628 [0.0911]
Reasoning	<b>Reasoning</b>	<b>Combo</b>	599	553	0.516 (0.0204)	0.495 (0.0213)	0.020 (0.0295)	0.9650 [0.4899]
Reasoning	<b>Emotions</b>	<b>Combo</b>	587	553	0.467 (0.0206)	0.495 (0.0213)	-0.029 (0.0296)	0.9046 [0.3329]
Emotions	<b>Emotions</b>	<b>Reasoning</b>	585	594	0.419 (0.0204)	0.498 (0.0205)	-0.080 (0.0290)	0.0563 [0.0031]
Emotions	<b>Emotions</b>	<b>Combo</b>	585	587	0.419 (0.0204)	0.489 (0.0206)	-0.070 (0.0290)	0.1248 [0.0079]
Emotions	<b>Reasoning</b>	<b>Combo</b>	594	587	0.498 (0.0205)	0.489 (0.0206)	0.009 (0.0291)	0.9650 [0.7472]
Combo	<b>Combo</b>	<b>Reasoning</b>	587	594	0.417 (0.0208)	0.416 (0.0204)	0.001 (0.0292)	0.9650 [0.9765]
Combo	<b>Combo</b>	<b>Emotions</b>	564	579	0.417 (0.0208)	0.399 (0.0204)	0.018 (0.0291)	0.9650 [0.5430]
Combo	<b>Reasoning</b>	<b>Emotions</b>	582	579	0.416 (0.0204)	0.399 (0.0204)	0.017 (0.0289)	0.9650 [0.5596]

Table A5: Average Treatment Effects: Misinformation Sharing by type of post - Follow-up

*Notes:* Sample includes the 2,554 participants in the Reasoning, Emotions or Combo intervention assignment groups who completed the follow-up survey and shared at least one non-misinformation post in the post survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Sharing Rate for misinformation posts, as defined in 1 -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-, and subsetting to the post type indicated in the first column of the table. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 4 tests using 10000 simulations and therefore have a minimum of 0.0001.

## A.5 Family 5 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>Facts</b>	3,283	1,069	0.348 (0.0083)	0.318 (0.0142)	0.030 (0.0165)	0.1942 [0.0355]
<b>Reasoning</b>	<b>Facts</b>	1,072	1,069	0.340 (0.0145)	0.318 (0.0142)	0.021 (0.0203)	0.5966 [0.1450]
<b>Emotions</b>	<b>Facts</b>	1,122	1,069	0.364 (0.0144)	0.318 (0.0142)	0.046 (0.0202)	0.0770 [0.0122]
<b>Combo</b>	<b>Facts</b>	1,089	1,069	0.340 (0.0144)	0.318 (0.0142)	0.022 (0.0202)	0.5966 [0.1417]
<b>Facts</b>	<b>No-course</b>	1,069	964	0.318 (0.0142)	0.318 (0.0150)	-0.000 (0.0207)	0.9845 [0.9842]

Table A6: Self-reported Social Media Behavior (Q4) in Follow-up

*Notes:* Sample includes 5,316 the participants who completed the follow-up survey. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the proportion of participants whose response to Question 4 in the follow-up survey contained one of the keywords. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 5 tests using 10000 simulations and therefore have a minimum of 0.0001.

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>Facts</b>	3,283	1,069	0.334 (0.0081)	0.275 (0.0140)	0.0015 (0.0160)	0.0010 [0.0001]
<b>Reasoning</b>	<b>Facts</b>	1,072	1,069	0.326 (0.0139)	0.275 (0.0140)	0.051 (0.0198)	0.0402 [0.0047]
<b>Emotions</b>	<b>Facts</b>	1,122	1,069	0.350 (0.0140)	0.275 (0.0140)	0.075 (0.0197)	0.0014 [0.0001]
<b>Combo</b>	<b>Facts</b>	1,089	1,069	0.325 (0.0142)	0.275 (0.0140)	0.050 (0.0197)	0.0416 [0.0056]
<b>Facts</b>	<b>No-course</b>	1,069	964	0.275 (0.0140)	0.293 (0.0142)	-0.018 (0.0200)	0.6145 [0.3824]

Table A7: Self-reported Tip to Friend (Q5) in Follow-up

*Notes:* Sample includes the 5,316 participants who completed the follow-up survey. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the proportion of participants whose response to Question 5 in the follow-up survey contained one of the keywords. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 5 tests using 10000 simulations and therefore have a minimum of 0.0001.

## A.6 Family 6 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Pre-survey</b>							
<b>Acc. Inter</b>	<b>Acc. After</b>	4,362	4,322	0.512 (0.0052)	0.580 (0.0047)	-0.067 (0.0071)	0.0001 [0.0000]
<b>Post-survey</b>							
<b>Acc. Inter, No-course</b>	<b>Acc. After, No-course</b>	801	827	0.513 (0.0122)	0.578 (0.0111)	-0.064 (0.0165)	0.0003 [0.0000]
<b>Acc. After, Treat. courses</b>	<b>Acc. After, No-course</b>	2,609	827	0.398 (0.0064)	0.578 (0.0111)	-0.179 (0.0128)	0.0001 [0.0000]
<b>Acc. Inter, Treat. courses</b>	<b>Acc. Inter, No-course</b>	2,657	801	0.353 (0.0064)	0.513 (0.0122)	-0.160 (0.0138)	0.0001 [0.0000]
<b>Acc. Inter, Treat. courses</b>	<b>Acc. After, Treat. courses</b>	2657	2609	0.353 (0.0064)	0.398 (0.0064)	-0.045 (0.0091)	0.0001 [0.0000]
<b>Difference-in-Differences</b>						0.019 (0.0188)	0.3097 [0.3093]

Table A8: Effect of Accuracy Nudge and Treatment Courses: Post-Survey Misinformation Sharing

*Notes:* Sample includes the 8,768 participants who completed the post-survey. All rows but the last one displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the proportion of misinformation posts shared in the pre- or post-survey, as indicated. The last row displays the difference of the third and fourth differences or the second and fifth differences. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 6 tests using 10000 simulations and therefore have a minimum of 0.0001.

## B Alternative outcome measures tests

### B.1 Family 7 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>No-course</b>	5,266	1,628	0.513 (0.0265)	0.041 (0.0465)	0.472 (0.0535)	0.0001 [0.0000]
<b>Treatment courses</b>	<b>Facts</b>	5,266	1,790	0.513 (0.0265)	0.215 (0.0438)	0.298 (0.0512)	0.0001 [0.0000]
<b>Reasoning</b>	<b>No-course</b>	1,690	1,628	0.355 (0.0467)	0.041 (0.0465)	0.314 (0.0659)	0.0001 [0.0000]
<b>Emotions</b>	<b>No-course</b>	1,836	1,628	0.617 (0.0447)	0.041 (0.0465)	0.577 (0.0645)	0.0001 [0.0000]
<b>Combo</b>	<b>No-course</b>	1,740	1,628	0.556 (0.0462)	0.041 (0.0465)	0.515 (0.0656)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	1,690	1,790	0.355 (0.0467)	0.215 (0.0438)	0.141 (0.0640)	0.0435 [0.0141]
<b>Emotions</b>	<b>Facts</b>	1,836	1,790	0.617 (0.0447)	0.215 (0.0438)	0.403 (0.0626)	0.0001 [0.0000]
<b>Combo</b>	<b>Facts</b>	1,740	1,790	0.556 (0.0462)	0.215 (0.0438)	0.341 (0.0637)	0.0001 [0.0000]
<b>Facts</b>	<b>No-course</b>	1,790	1,628	0.215 (0.0438)	0.041 (0.0465)	0.174 (0.0639)	0.0141 [0.0032]
<b>Reasoning</b>	<b>Emotions</b>	1,690	1,836	0.355 (0.0467)	0.617 (0.0447)	-0.262 (0.0647)	0.0002 [0.0001]
<b>Reasoning</b>	<b>Combo</b>	1,690	1,740	0.355 (0.0467)	0.556 (0.0462)	-0.201 (0.0657)	0.0002 [0.0023]
<b>Emotions</b>	<b>Combo</b>	1,836	1,740	0.617 (0.0447)	0.556 (0.0462)	0.061 (0.0643)	0.3419 [0.3404]

Table B1: Average Treatment Effects - Change in Sharing Discernment Score

*Notes:* Sample includes the 8,768 participants who completed the post-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first column of the table, the t-tests compare the post-pre difference in the discernment score, defined as the number of non-misinformation posts shared minus the number of misinformation posts shared. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 7 tests using 10000 simulations and therefore have a minimum of 0.0001.

## B.2 Family 8 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>No-course</b>	5,266	1,628	-0.182 (0.0051)	-0.027 (0.0084)	-0.155 (0.0099)	0.0001 [0.0000]
<b>Treatment courses</b>	<b>Facts</b>	5,266	1,790	-0.182 (0.0051)	-0.082 (0.0083)	-0.099 (0.0097)	0.0001 [0.0000]
<b>Reasoning</b>	<b>No-course</b>	1,690	1,628	-0.150 (0.0089)	-0.027 (0.0084)	-0.124 (0.0123)	0.0001 [0.0000]
<b>Emotions</b>	<b>No-course</b>	1,836	1,628	-0.204 (0.0089)	-0.027 (0.0084)	-0.177 (0.0122)	0.0001 [0.0000]
<b>Combo</b>	<b>No-course</b>	1,740	1,628	-0.189 (0.0089)	-0.027 (0.0084)	-0.162 (0.0123)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	1,690	1,790	-0.150 (0.0089)	-0.082 (0.0083)	-0.068 (0.0121)	0.0001 [0.0000]
<b>Emotions</b>	<b>Facts</b>	1,836	1,790	-0.204 (0.0089)	-0.082 (0.0083)	-0.121 (0.0121)	0.0001 [0.0000]
<b>Combo</b>	<b>Facts</b>	1,740	1,790	-0.189 (0.0089)	-0.082 (0.0083)	-0.106 (0.0122)	0.0001 [0.0000]
<b>Facts</b>	<b>No-course</b>	1,790	1,628	-0.082 (0.0083)	-0.027 (0.0084)	-0.056 (0.0118)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Emotions</b>	1,690	1,836	-0.150 (0.0089)	-0.204 (0.0089)	0.053 (0.0126)	0.0003 [0.0000]
<b>Reasoning</b>	<b>Combo</b>	1,690	1,740	-0.150 (0.0089)	-0.189 (0.0089)	0.038 (0.0126)	0.0053 [0.0025]
<b>Emotions</b>	<b>Combo</b>	1,836	1,740	-0.204 (0.0089)	-0.189 (0.0089)	-0.015 (0.0126)	0.2260 [0.2214]

Table B2: Average Treatment Effects - Non-Corresponding Misinformation Sharing

*Notes:* Sample includes the 8,768 participants who completed the post-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first column of the table, the t-tests compare the post-pre difference in the proportion of misinformation posts shared, excluding those misinformation posts in the post-survey that are counterparts of non-misinformation posts in the pre-survey. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 8 tests using 10000 simulations and therefore have a minimum of 0.0001.



### B.3 Family 9 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>No-course</b>	2,831	863	0.201 (0.0070)	0.295 (0.0145)	-0.094 (0.0161)	0.0001 [0.0000]
<b>Treatment courses</b>	<b>Facts</b>	2,831	1,008	0.201 (0.0070)	0.233 (0.0124)	-0.032 (0.0142)	0.0819 [0.0117]
<b>Reasoning</b>	<b>No-course</b>	890	863	0.213 (0.0129)	0.295 (0.0145)	-0.082 (0.0194)	0.0004 [0.0000]
<b>Emotions</b>	<b>No-course</b>	996	863	0.195 (0.0116)	0.295 (0.0145)	-0.099 (0.0185)	0.0001 [0.0000]
<b>Combo</b>	<b>No-course</b>	945	863	0.196 (0.0119)	0.295 (0.0145)	-0.099 (0.0187)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	890	1,008	0.213 (0.0129)	0.233 (0.0124)	-0.020 (0.0179)	0.5434 [0.1302]
<b>Emotions</b>	<b>Facts</b>	996	1,008	0.195 (0.0116)	0.233 (0.0124)	-0.038 (0.0170)	0.0852 [0.0125]
<b>Combo</b>	<b>Facts</b>	945	1,008	0.196 (0.0119)	0.233 (0.0124)	-0.038 (0.0172)	0.0865 [0.0142]
<b>Facts</b>	<b>No-course</b>	1,008	863	0.233 (0.0124)	0.295 (0.0145)	-0.061 (0.0191)	0.0080 [0.0006]
<b>Reasoning</b>	<b>Emotions</b>	890	996	0.213 (0.0129)	0.195 (0.0116)	0.018 (0.0173)	0.5558 [0.3033]
<b>Reasoning</b>	<b>Combo</b>	890	945	0.213 (0.0129)	0.196 (0.0119)	0.018 (0.0176)	0.5558 [0.3184]
<b>Emotions</b>	<b>Combo</b>	996	945	0.195 (0.0116)	0.196 (0.0119)	-0.000 (0.0166)	0.9836 [0.9847]

Table B3: Average Treatment Effects: Opposite Outcome

*Notes:* Sample includes the 7,688 participants who completed the post-survey and shared less than all three non-misinformation posts in the pre-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Opposite Sharing Rate for misinformation posts, as defined in 2. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 9 tests using 10000 simulations and therefore have a minimum of 0.0001.

## B.4 Family 10 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Treatment courses</b>	<b>No-course</b>	5,266	1,628	1.467 (0.1173)	-0.157 (0.2152)	1.624 (0.2451)	0.0001 [0.0000]
<b>Treatment courses</b>	<b>Facts</b>	5,266	1,790	1.467 (0.1173)	0.880 (0.1998)	0.587 (0.2317)	0.0443 [0.0057]
<b>Reasoning</b>	<b>No-course</b>	1,690	1,628	1.157 (0.2099)	-0.157 (0.2152)	1.315 (0.3006)	0.0002 [0.0000]
<b>Emotions</b>	<b>No-course</b>	1,836	1,628	1.741 (0.1947)	-0.157 (0.2152)	1.898 (0.2902)	0.0001 [0.0000]
<b>Combo</b>	<b>No-course</b>	1,740	1,628	1.479 (0.2055)	-0.157 (0.2152)	1.637 (0.2975)	0.0001 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	1,690	1,790	1.157 (0.2099)	0.880 (0.1998)	0.277 (0.2898)	0.5045 [0.1696]
<b>Emotions</b>	<b>Facts</b>	1,836	1,790	1.741 (0.1947)	0.880 (0.1998)	0.860 (0.2790)	0.0092 [0.0010]
<b>Combo</b>	<b>Facts</b>	1,740	1,790	1.479 (0.2055)	0.880 (0.1998)	0.599 (0.2866)	0.1155 [0.0184]
<b>Facts</b>	<b>No-course</b>	1,790	1,628	0.880 (0.1998)	-0.157 (0.2152)	1.038 (0.2937)	0.0021 [0.0002]
<b>Reasoning</b>	<b>Emotions</b>	1,690	1,836	1.157 (0.2099)	1.741 (0.1947)	-0.583 (0.2862)	0.1175 [0.0416]
<b>Reasoning</b>	<b>Combo</b>	1,690	1,740	1.157 (0.2099)	1.479 (0.2055)	-0.322 (0.2937)	0.5045 [0.2731]
<b>Emotions</b>	<b>Combo</b>	1,836	1,740	1.741 (0.1947)	1.479 (0.2055)	0.261 (0.2830)	0.5045 [0.3557]

Table B4: Average Treatment Effects: Change in Accuracy Discernment Score

*Notes:* Sample includes the 8,768 participants who completed the post-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first column of the table, the t-tests compare the post-pre difference in the accuracy discernment score. Accuracy scores are defined using a numerical encoding of the answer to the “To the best of your knowledge, how accurate is the claim in the above post?” where “Not at all accurate” is encoded as  $-3$  or  $3$  if the post is a non-misinformation post or a misinformation post, respectively; “Not very accurate” is encoded as  $-1$  or  $1$  if the post is a non-misinformation post or a misinformation post, respectively; “Somewhat accurate” is encoded as  $1$  or  $-1$  if the post is a non-misinformation post or a misinformation post, respectively; and “Very accurate” is encoded as  $3$  or  $-3$  if the post is a non-misinformation post or a misinformation post, respectively. Summing this numerical encoding of all posts a participant saw in either the pre- or the post-survey yields the accuracy discernment score. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 10 tests using 10000 simulations and therefore have a minimum of 0.0001.

## B.5 Family 11 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Pre-survey</b>							
<b>Acc. Inter</b>	<b>Acc. After</b>	4,362	4,322	-1.142 (0.0243)	-1.251 (0.0239)	0.109 (0.0341)	0.0038 [0.0007]
<b>Post-survey</b>							
<b>Acc. Inter, No-course</b>	<b>Acc. After, No-course</b>	801	827	-1.164 (0.0555)	-1.308 (0.0531)	0.145 (0.0768)	0.1276 [0.0298]
<b>Acc. After, Treat. courses</b>	<b>Acc. After, No-course</b>	2,609	827	-0.686 (0.0309)	-1.308 (0.0531)	0.623 (0.0615)	0.0001 [0.0000]
<b>Acc. Inter, Treat. courses</b>	<b>Acc. Inter, No-course</b>	2,657	801	-0.639 (0.0292)	-1.164 (0.0555)	0.524 (0.0627)	0.0001 [0.0000]
<b>Acc. Inter, Treat. courses</b>	<b>Acc. After, Treat. courses</b>	2,657	2,609	-0.639 (0.0292)	-0.686 (0.0309)	0.046 (0.0425)	0.4598 [0.2764]
<b>Difference-in-Differences</b>						0.019 (0.0188)	0.4598 [0.3093]

Table B5: Effect of Accuracy Nudge and Treatment Courses: Post-Survey Sharing Discernment Score

*Notes:* Sample includes the 8,768 participants who completed the post-survey. All rows but the last one displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the pre- or post-survey sharing discernment score as indicated, defined as the number of non-misinformation posts shared minus the number of misinformation posts shared. The last row displays the difference of the third and fourth differences or the second and fifth differences. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 11 tests using 10000 simulations and therefore have a minimum of 0.0001.

## B.6 Family 12 Tests

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Pre-survey</b>							
<b>Acc. Inter</b>	<b>Acc. After</b>	4,362	4,322	2.316 (0.1072)	1.558 (0.1044)	0.758 (0.1496)	0.0001 [0.0000]
<b>Post-survey</b>							
<b>Acc. Inter, No-course</b>	<b>Acc. After, No-course</b>	801	827	1.779 (0.2362)	1.063 (0.2336)	0.716 (0.3322)	0.0415 [0.0156]
<b>Acc. After, Treat. courses</b>	<b>Acc. After, No-course</b>	2,609	827	3.146 (0.1278)	1.063 (0.2336)	2.083 (0.2663)	0.0001 [0.0000]
<b>Acc. Inter, Treat. courses</b>	<b>Acc. Inter, No-course</b>	2,657	801	3.896 (0.1278)	1.779 (0.2362)	2.117 (0.2686)	0.0001 [0.0000]
<b>Acc. Inter, Treat. courses</b>	<b>Acc. After, Treat. courses</b>	2,657	2,609	3.896 (0.1278)	3.146 (0.1278)	0.751 (0.1807)	0.0001 [0.0000]
<b>Difference-in-Differences</b>						0.035 (0.3782)	0.9307 [0.9269]

Table B6: Effect of Accuracy Nudge and Treatment Courses: Post-Survey Accuracy Discernment Score

*Notes:* Sample includes the 8,768 participants who completed the post-survey. All rows but the last one displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the pre- pr post-survey accuracy discernment score as indicated. Accuracy scores are defined using a numerical encoding of the answer to the “To the best of your knowledge, how accurate is the claim in the above post?” where “Not at all accurate” is encoded as  $-3$  or  $3$  if the post is a non-misinformation post or a misinformation post, respectively; “Not very accurate” is encoded as  $-1$  or  $1$  if the post is a non-misinformation post or a misinformation post, respectively; “Somewhat accurate” is encoded as  $1$  or  $-1$  if the post is a non-misinformation post or a misinformation post, respectively; and “Very accurate” is encoded as  $3$  or  $-3$  if the post is a non-misinformation post or a misinformation post, respectively. Summing this numerical encoding of all posts a participant saw in the post-survey yields the accuracy discernment score. The last row displays the difference of the third and fourth differences or the second and fifth differences. Standard errors between parentheses. Unadjusted p-Values between brackets. Romano-Wolf adjusted p-values are computed jointly for all Family 12 tests using 1000 simulations and therefore have a minimum of 0.001.

## C Additional Tables and Figures

### C.1 Ad Cost and Text Message Course Funnel Statistics

	Count	% of Previous Funnel Stage	Cost Per
<b>Impressions</b>	19,640,962	-	\$0.001
<b>Clicks</b>	193,227	0.98%	\$0.054
<b>Started Pre-survey</b>	25,287	13.08%	\$0.41
<b>Completed Pre-survey</b>	22,526	89.08%	\$0.46
<b>Started Text Message Course</b>	18,598	82.56%	\$0.56
<b>Completed Day 1 Course</b>	16,684	89.71%	\$0.63
<b>Completed Day 2 Course</b>	13,997	83.89%	\$0.75
<b>Completed Day 3 Course</b>	13,093	93.54%	\$0.80
<b>Completed Day 4 Course</b>	11,396	87.04%	\$0.92
<b>Completed Entire Course</b>	10,934	95.95%	\$0.96
<b>Started Post-survey</b>	9,589	87.70%	\$1.09
<b>Completed Post-survey</b>	8,684	90.56%	\$1.20
<b>Started Follow-up</b>	5,785	66.62%	\$1.81
<b>Completed Follow-up</b>	5,316	91.88%	\$1.97

Table C1: Full Funnel with Ad Costs

*Notes:* The ads metrics, Impressions and Clicks, are extracted from Facebook Ads manager. The total cost of ads was \$10,460.26. The number of participants who started the pre-survey is an upper bound estimate because we could count only survey copies on Qualtrics and not identify users. We were only able to identify users once they completed the pre-survey and provided a phone number, which is how we ensured that all down-funnel outcomes counted unique users based on phone number. The post- and follow-up surveys required users to validate their phone number before they could start the survey. We received 40,845 survey copies in total for the pre-survey, of which we discarded 3,092 users who had participated in one of our pilot studies and another 12,466 survey copies filled out by duplicated phone numbers. In the post-survey, we filtered out participants who encountered system errors ( $N = 104$ ), did not have at least 5 days in between pre- and post-survey dates ( $N = 2,101$ ), and/or did not complete the full text-message course ( $N = 509$ ). For the follow-up survey, we filtered out 2,369 survey copies filled out by duplicated phone numbers and retained only the first copy done by each phone number.

	No-course baseline	Facts baseline	Reasoning course	Emotions course	Combo course
<b>Started Text Message Course</b>	1,646	3,746	3,686	3,755	3,713
	-	-	-	-	-
<b>Completed Day 1 Course</b>	1,353 (82.2%)	3,276 (87.45%)	3,234 (87.74%)	3,240 (86.28%)	3,236 (87.15%)
<b>Completed Day 2 Course</b>	1,122 (68.17%)	2,592 (69.19%)	2,575 (69.86%)	2,551 (67.94%)	2,581 (69.51%)
<b>Completed Day 3 Course</b>	1,029 (62.52%)	2,356 (62.89%)	2,260 (61.31%)	2,373 (63.2%)	2,407 (64.83%)
<b>Completed Day 4 Course</b>	927 (56.32%)	2,195 (58.6%)	2,060 (55.89%)	2,225 (59.25%)	2,194 (59.09%)
<b>Completed Entire Course</b>	856 (52%)	2,090 (55.79%)	1,960 (53.17%)	2,124 (56.56%)	2,051 (55.24%)

Table C2: Text Message Course Funnel, by Assignment Group

*Notes:* The numbers in percent (in brackets) represent the % of participants who started the text message course in the respective assignment groups.

## C.2 Covariates Explanation

Covariate in Survey	Constructed Covariate	Explanation
<b>Age</b>	Age	Integer denoting age of participants
<b>Gender</b>	Man	Complement to "woman" and "other"
<b>Education</b>	High school or less	including "Less than a high school diploma" and "High school degree or equivalent"
	Some college	including "Some college, no degree" and "Associate degree"
	Bachelor's degree	Bachelor's degree
	Graduate degrees	including "Master's degree" and "Doctorate or professional degree"
<b>Marital status</b>	Married	Complement to "Single," "Widowed," "Divorced," and "Separated"
<b>Employment status</b>	Employed	including "Employed full time," "Employed part time," and "Self-employed"
	Unemployed	including "Unemployed and currently looking for work," "Unemployed not currently looking for work," "Retired," "Homemaker," and "Unable to work"
<b>Location</b>	Student	Student
	Mostly urban	Live in mostly urban area
	Suburban	live in suburban area
<b>Religion</b>	Mostly rural	live in mostly rural area
	Christian	Complement to "None," "Hinduism," "Muslim," "Traditionalist," and "Other"
<b>Religiosity</b>	Attends religious services	Frequency of attending religious services including "Less than once a month," "One to three times per month," "Once a week," "More than once a week but less than daily," and "Daily"; Complement to "Never"
<b>Social media user</b>	Uses social media	1 if "Yes" and 0 if "No"
<b>Hrs/day on social media</b>	Hrs/day on social media	Integer denoting how many hours are spent on average each day on social media
<b>Prop. of content shared</b>	80 - 100%	Proportion of the content the participant sees on social media that they choose to share is between 80 and 100%
	60 - 80%	Similar as above
	40 - 60%	Similar as above
	20 - 40%	Similar as above
	0 - 20%	Similar as above

Table C3: Covariates Explanation

### C.3 Attrition and Balance Figures

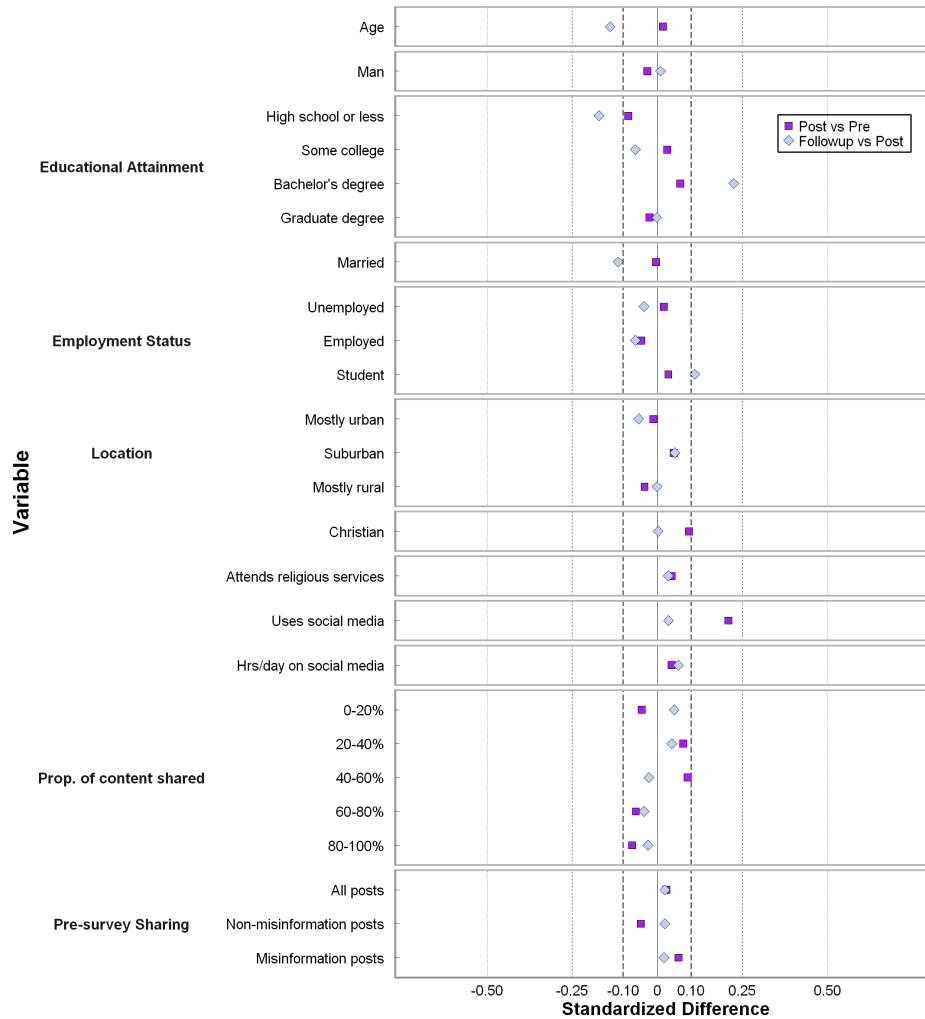


Figure C1: Differences in Samples, by Survey Completion

*Notes:* Pre sample includes the 13,842 participants who completed the pre-survey but not the post-survey. Post sample includes the 3,368 participants who completed both the pre-survey and the post-survey, but not the follow-up. Followup sample includes the 5,316 participants who completed all of the surveys. The differences are taken by subtracting the Pre sample from the Post sample, and the Post sample from the Follow-up sample.



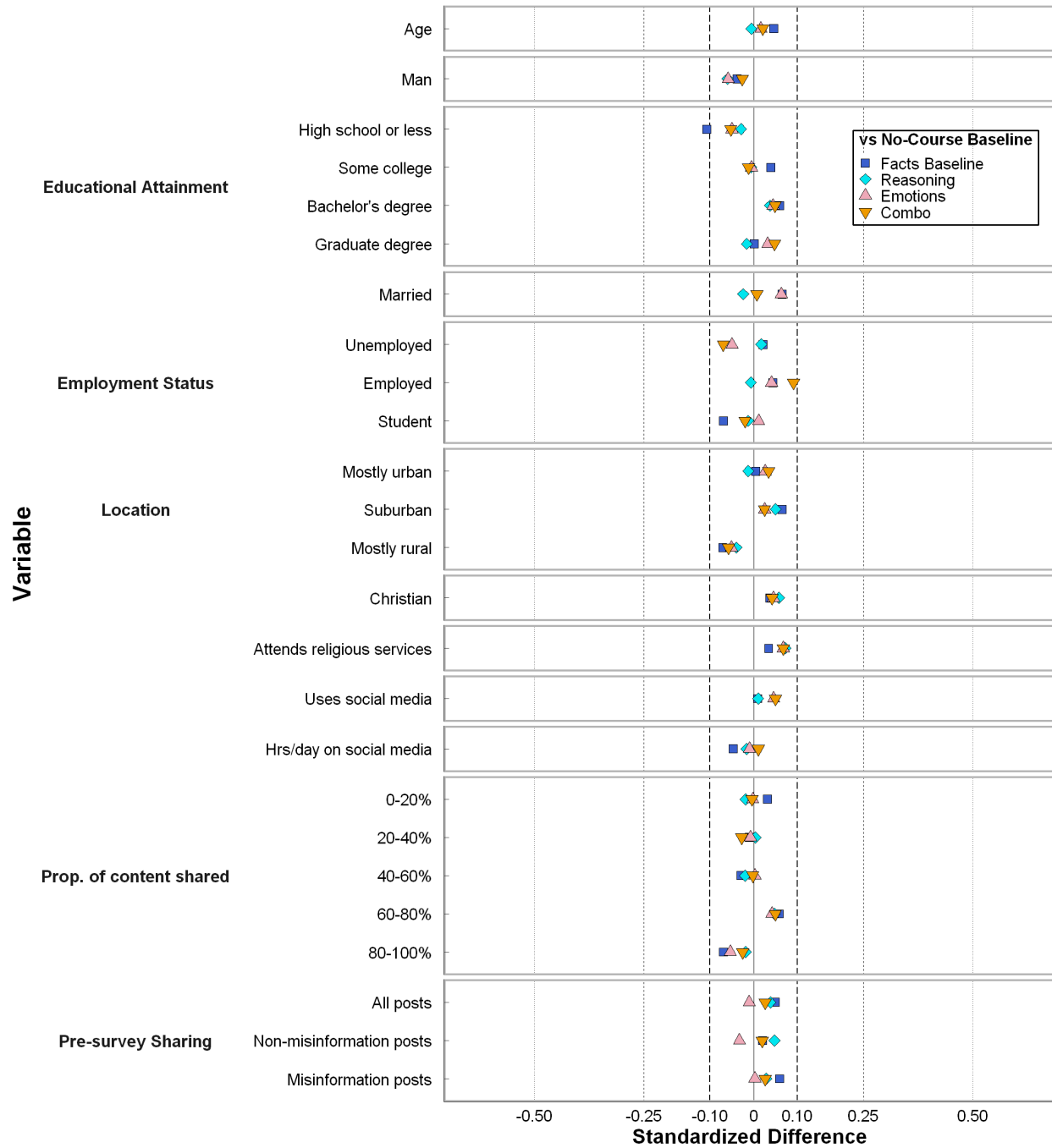


Figure C2: Difference in Samples of Post-survey Completers, by Intervention Assignment Group

Notes: Sample includes the 8,684 participants who completed the post-survey. The differences are taken by subtracting the No-course baseline from each of the other groups.

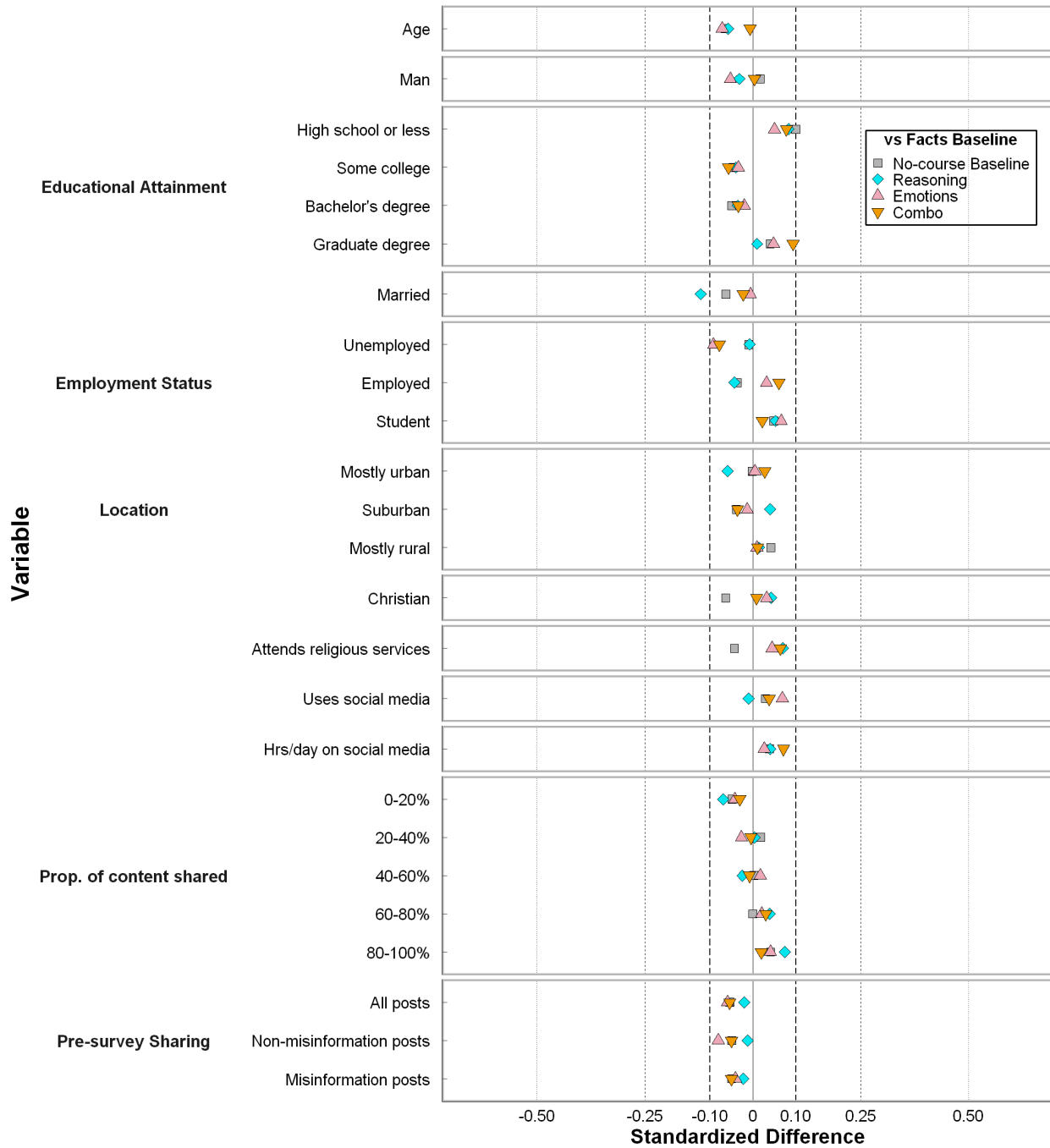


Figure C3: Difference in Samples of Follow-up Survey Completers, by Intervention Assignment Group

Notes: Sample includes the 5,316 participants who completed the post-survey. The differences are taken by subtracting the Facts baseline from each of the other groups.

## C.4 Main Results, by Accuracy Nudge Group

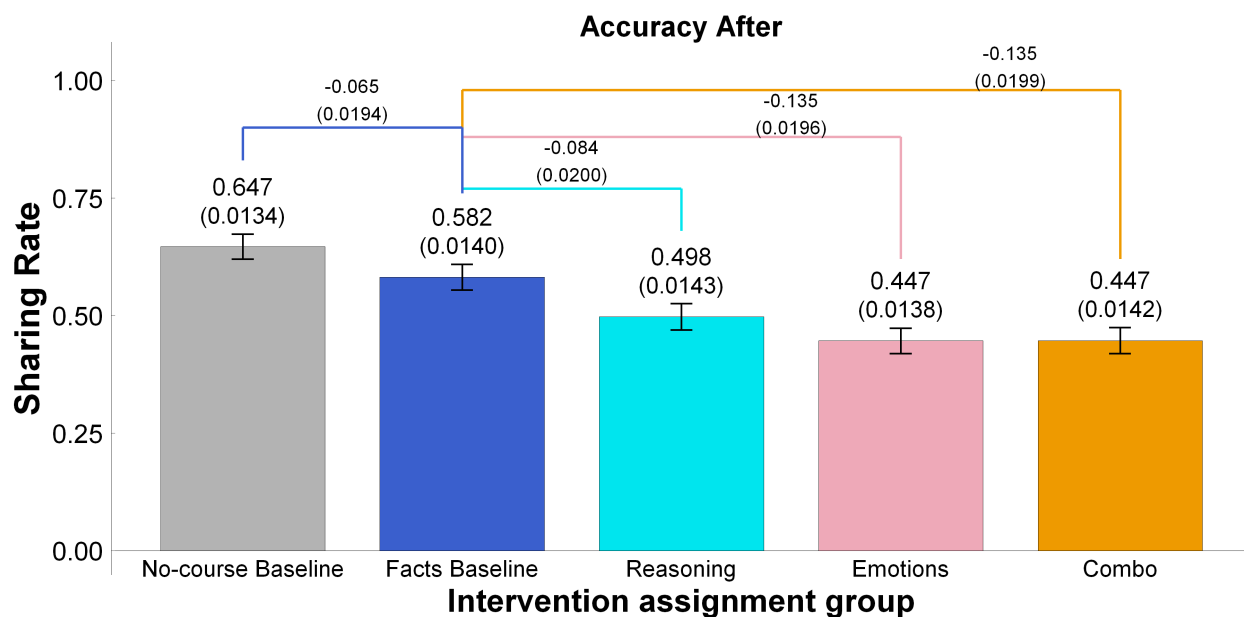


Figure C4: Misinformation Sharing, by Intervention Assignment Group

*Notes:* Sample includes the 3,984 participants from the Accuracy After group who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in Equation 1, by participants in their respective intervention assignment group. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

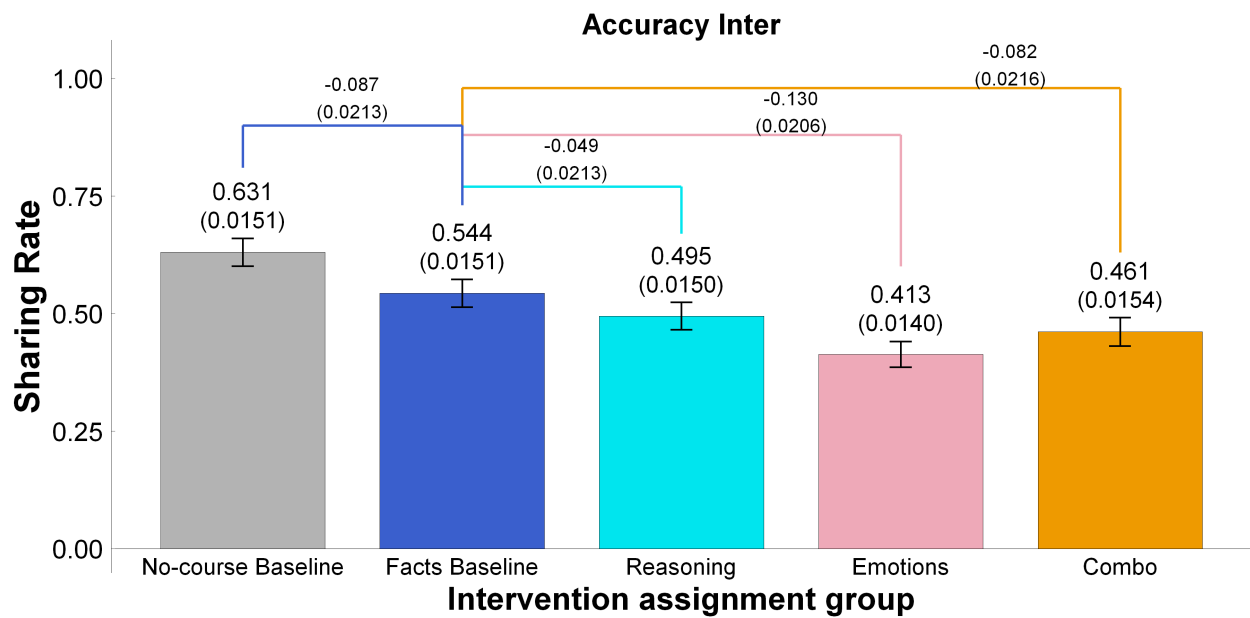


Figure C5: Misinformation Sharing, by Intervention Assignment Group

*Notes:* Sample includes the 3,704 participants from the Accuracy Inter group who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in Equation 1, by participants in their respective intervention assignment group. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

## C.5 Follow-up Results, by Prime Status

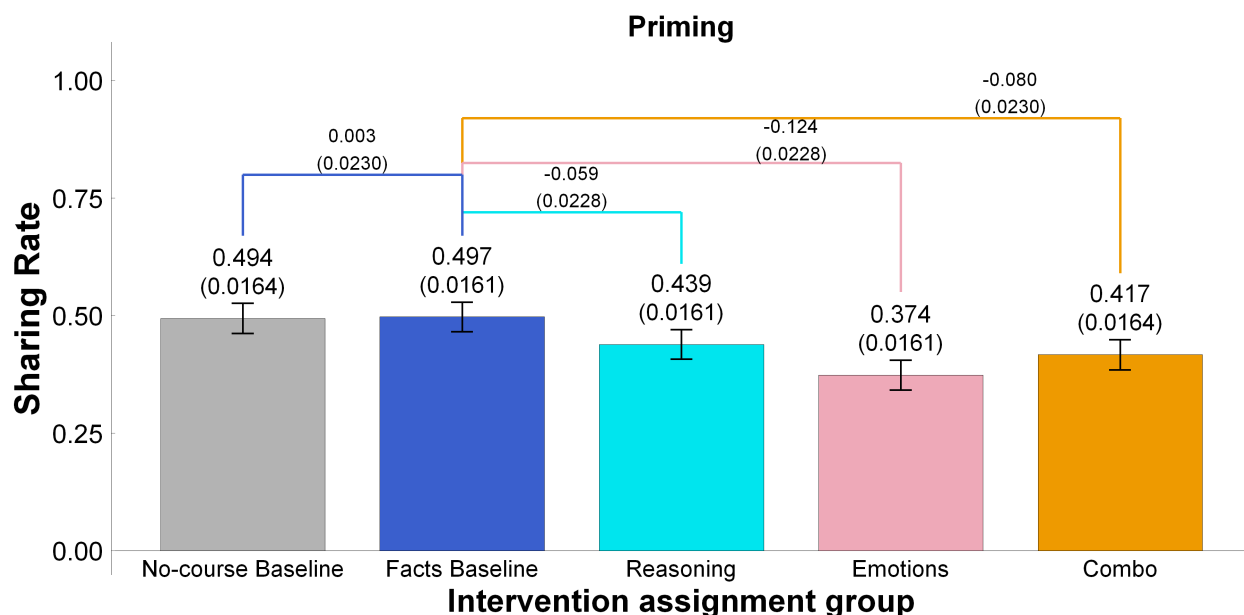


Figure C6: Misinformation Sharing in Follow-up, by Intervention Assignment Group - Primed Participants

*Notes:* Sample includes the 3,176 primed participants who completed the follow-up survey and shared at least one non-misinformation post in the post-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in 1, -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in the average Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

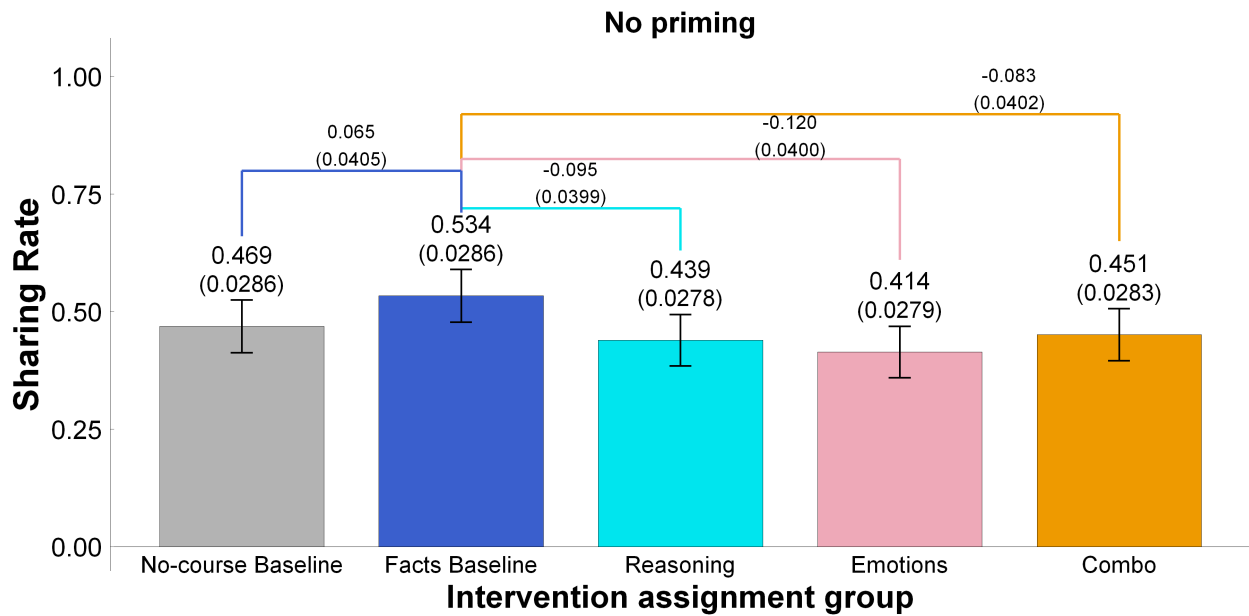


Figure C7: Misinformation Sharing in Follow-up, by Intervention Assignment Group - Non-Primed Participants

*Notes:* Sample includes the 1,075 non-primed participants who completed the follow-up survey and shared at least one non-misinformation post in the post-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in 1, -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in the average Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

## D Robustness to sample who correctly answered all attention check questions

	No-course baseline	Facts baseline	Reasoning course	Emotions course	Combo course	Treatment courses	Baselines	All
<b>Number of Observations</b>								
Accuracy After	256	292	273	299	300	872	548	1,420
Accuracy Inter	300	373	341	384	358	1,083	673	1,756
All	556	665	614	683	658	1,955	1,221	3,176
<b>Pre Non-misinfo Posts</b>								
Accuracy After	0.637	0.605	0.646	0.615	0.595	0.618	0.619	0.618
	(0.0209)	(0.0183)	(0.0190)	(0.0181)	(0.0201)	(0.0110)	(0.0138)	(0.0086)
Accuracy Inter	0.736	0.710	0.722	0.717	0.664	0.700	0.722	0.709
	(0.0200)	(0.0191)	(0.0209)	(0.0205)	(0.0206)	(0.0120)	(0.0138)	(0.0091)
All	0.682	0.651	0.680	0.659	0.627	0.655	0.665	0.659
	(0.0147)	(0.0134)	(0.0141)	(0.0137)	(0.0145)	(0.0082)	(0.0099)	(0.0063)
<b>Pre Misinfo Posts</b>								
Accuracy After	0.488	0.449	0.480	0.467	0.460	0.469	0.467	0.468
	(0.0196)	(0.0170)	(0.0185)	(0.0167)	(0.0177)	(0.0102)	(0.0129)	(0.0080)
Accuracy Inter	0.563	0.567	0.548	0.534	0.509	0.530	0.565	0.544
	(0.0192)	(0.0182)	(0.0205)	(0.0191)	(0.0187)	(0.0112)	(0.0132)	(0.0086)
All	0.523	0.501	0.510	0.497	0.483	0.496	0.511	0.502
	(0.0139)	(0.0126)	(0.0138)	(0.0126)	(0.0129)	(0.0076)	(0.0093)	(0.0059)
<b>Post Non-misinfo Posts</b>								
Accuracy After	0.616	0.503	0.520	0.456	0.475	0.482	0.553	0.509
	(0.0215)	(0.0200)	(0.0203)	(0.0190)	(0.0201)	(0.0114)	(0.0148)	(0.0091)
Accuracy Inter	0.663	0.622	0.554	0.548	0.512	0.538	0.641	0.578
	(0.0225)	(0.0220)	(0.0226)	(0.0218)	(0.0224)	(0.0129)	(0.0158)	(0.0101)
All	0.637	0.555	0.535	0.496	0.492	0.507	0.593	0.540
	(0.0156)	(0.0150)	(0.0151)	(0.0144)	(0.0150)	(0.0086)	(0.0109)	(0.0068)
<b>Post Misinfo Posts</b>								
Accuracy After	0.464	0.366	0.368	0.283	0.314	0.320	0.410	0.354
	(0.0187)	(0.0177)	(0.0178)	(0.0150)	(0.0168)	(0.0096)	(0.0130)	(0.0078)
Accuracy Inter	0.523	0.478	0.421	0.365	0.381	0.388	0.499	0.431
	(0.0205)	(0.0204)	(0.0211)	(0.0188)	(0.0192)	(0.0114)	(0.0145)	(0.0091)
All	0.491	0.415	0.392	0.319	0.344	0.350	0.450	0.388
	(0.0139)	(0.0135)	(0.0137)	(0.0119)	(0.0127)	(0.0074)	(0.0098)	(0.0060)
<b>Primary Outcome: Sharing Rate</b>								
Accuracy After	0.631	0.544	0.495	0.413	0.461	0.455	0.584	0.505
	(0.0151)	(0.0151)	(0.0150)	(0.0140)	(0.0154)	(0.0085)	(0.0107)	(0.0068)
Accuracy Inter	0.647	0.582	0.498	0.447	0.447	0.464	0.614	0.524
	(0.0134)	(0.0140)	(0.0143)	(0.0138)	(0.0142)	(0.0081)	(0.0097)	(0.0064)
All	0.639	0.563	0.496	0.430	0.454	0.459	0.600	0.515
	(0.0100)	(0.0103)	(0.0103)	(0.0098)	(0.0105)	(0.0059)	(0.0072)	(0.0046)
<b>Pre Sharing Discernment</b>								
Accuracy After	-1.196	-1.100	-1.054	-1.188	-1.169	-1.140	-1.145	-1.142
	(0.0581)	(0.0525)	(0.0565)	(0.0526)	(0.0520)	(0.0310)	(0.0390)	(0.0243)
Accuracy Inter	-1.357	-1.269	-1.201	-1.167	-1.268	-1.211	-1.311	-1.251
	(0.0531)	(0.0513)	(0.0534)	(0.0547)	(0.0548)	(0.0314)	(0.0369)	(0.0239)
All	-1.278	-1.183	-1.128	-1.178	-1.218	-1.175	-1.228	-1.196
	(0.0393)	(0.0367)	(0.0389)	(0.0379)	(0.0377)	(0.0220)	(0.0269)	(0.0170)
<b>Post Sharing Discernment Score</b>								
Accuracy After	-1.164	-0.889	-0.728	-0.525	-0.678	-0.639	-1.018	-0.787
	(0.0555)	(0.0529)	(0.0529)	(0.0470)	(0.0519)	(0.0292)	(0.0384)	(0.0234)
Accuracy Inter	-1.308	-1.050	-0.818	-0.597	-0.646	-0.686	-1.175	-0.879
	(0.0531)	(0.0531)	(0.0556)	(0.0516)	(0.0533)	(0.0309)	(0.0377)	(0.0242)
All	-1.237	-0.969	-0.773	-0.560	-0.662	-0.662	-1.097	-0.833
	(0.0384)	(0.0375)	(0.0384)	(0.0348)	(0.0371)	(0.0212)	(0.0269)	(0.0168)
<b>Pre Accuracy Discernment Score</b>								
Accuracy After	1.924	2.394	2.803	2.176	2.283	2.408	2.173	2.316
	(0.2555)	(0.2323)	(0.2489)	(0.2271)	(0.2373)	(0.1371)	(0.1720)	(0.1072)
Accuracy Inter	1.232	1.445	1.451	1.843	1.799	1.700	1.342	1.558
	(0.2402)	(0.2231)	(0.2315)	(0.2348)	(0.2373)	(0.1354)	(0.1635)	(0.1044)
All	1.572	1.924	2.118	2.014	2.044	2.057	1.757	1.939
	(0.1753)	(0.1614)	(0.1706)	(0.1633)	(0.1678)	(0.0965)	(0.1189)	(0.0749)
<b>Post Accuracy Discernment Score</b>								
Accuracy After	1.779	3.319	3.645	4.142	3.872	3.896	2.595	3.388
	(0.2362)	(0.2326)	(0.2292)	(0.2089)	(0.2273)	(0.1278)	(0.1669)	(0.1020)
Accuracy Inter	1.063	2.280	2.916	3.347	3.165	3.146	1.692	2.570
	(0.2336)	(0.2226)	(0.2269)	(0.2142)	(0.2232)	(0.1278)	(0.1618)	(0.1009)
All	1.415	2.804	3.276	3.755	3.523	3.524	2.143	2.981
	(0.1663)	(0.1615)	(0.1614)	(0.1498)	(0.1595)	(0.0905)	(0.1164)	(0.0719)

Table D1: Summary of Outcomes by Text Message Course Intervention and Accuracy Nudge Assignment - Attention Check Sample



*Notes:* Sample includes 3,172 participants who completed the post-survey, except for the “Primary Outcome: Sharing Rate” rows that exclude 434 participants who did not share at least one non-misinformation post in the pre-survey. The first two rows display the number of observations in each assignment group. The other rows display averages by assignment group, with standard errors in parentheses.

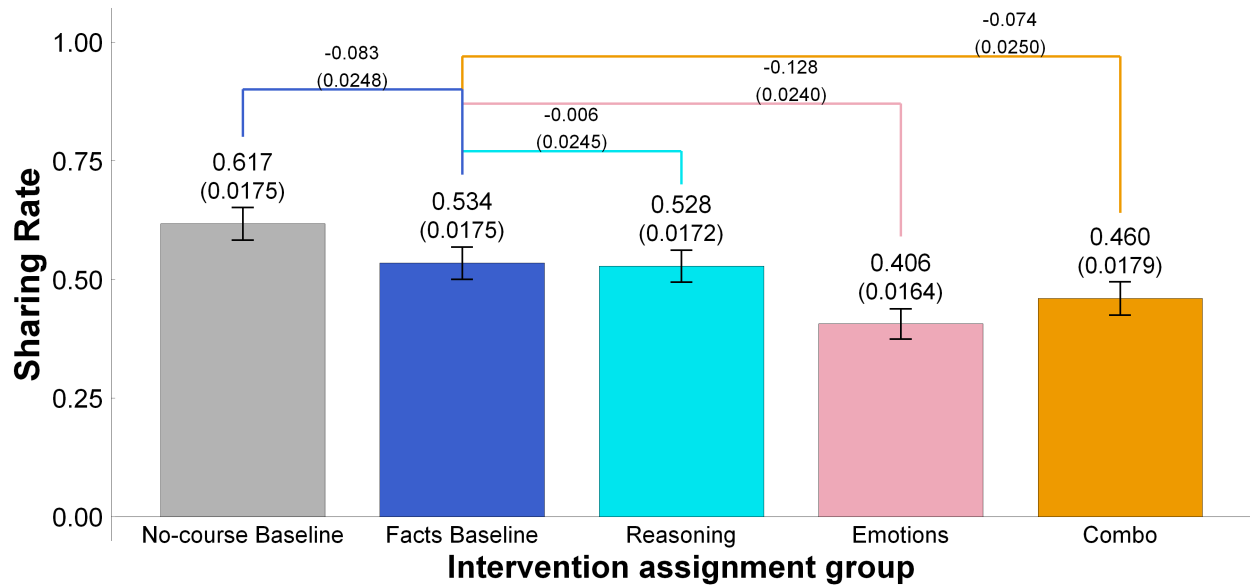


Figure D1: Misinformation Sharing, by Intervention Assignment Group - Attention Check Sample

*Notes:* Sample includes the 2,741 participants who completed the post-survey passing attention checks in the pre- and post- survey, and shared at least one non-misinformation post in the pre-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in Equation 1, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

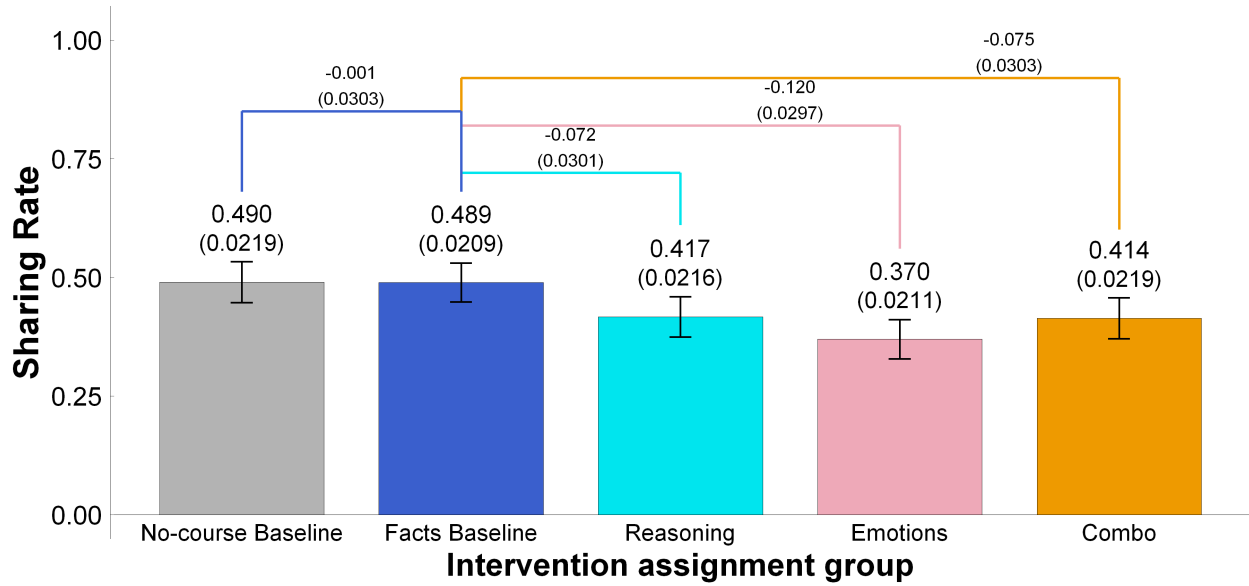


Figure D2: Misinformation Sharing, by Intervention Assignment Group - Followup - Attention Check Sample

*Notes:* Sample includes the 1,798 participants who completed the follow-up survey passing all the attention checks and shared at least one non-misinformation post in the post-survey. Each bar displays the Sharing Rate for misinformation posts, as defined in 1, -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in the average Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

## E Subgroup Analysis

### E.1 Main Results, by Gender

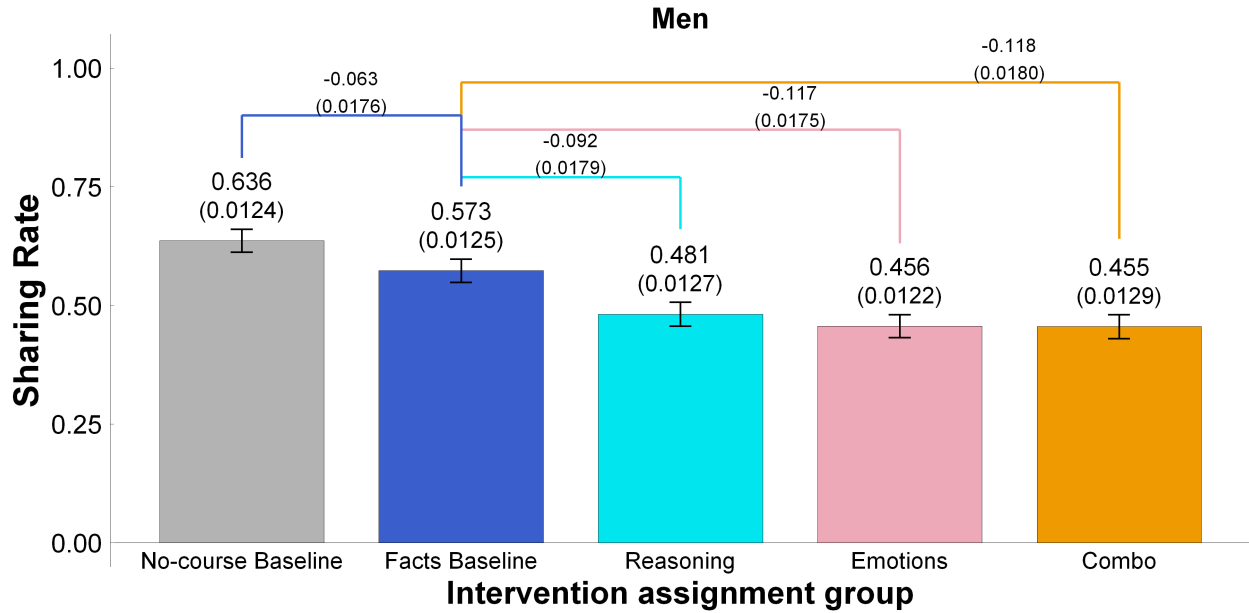


Figure E1: Primary outcome by treatment - Men

*Notes:* Sample includes the 3,111 participants who were assigned to the Reasoning or Emotions courses and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each bar displays the Sharing Rate for misinformation posts, as defined in Equation 1, by participants in their respective intervention assignment group. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

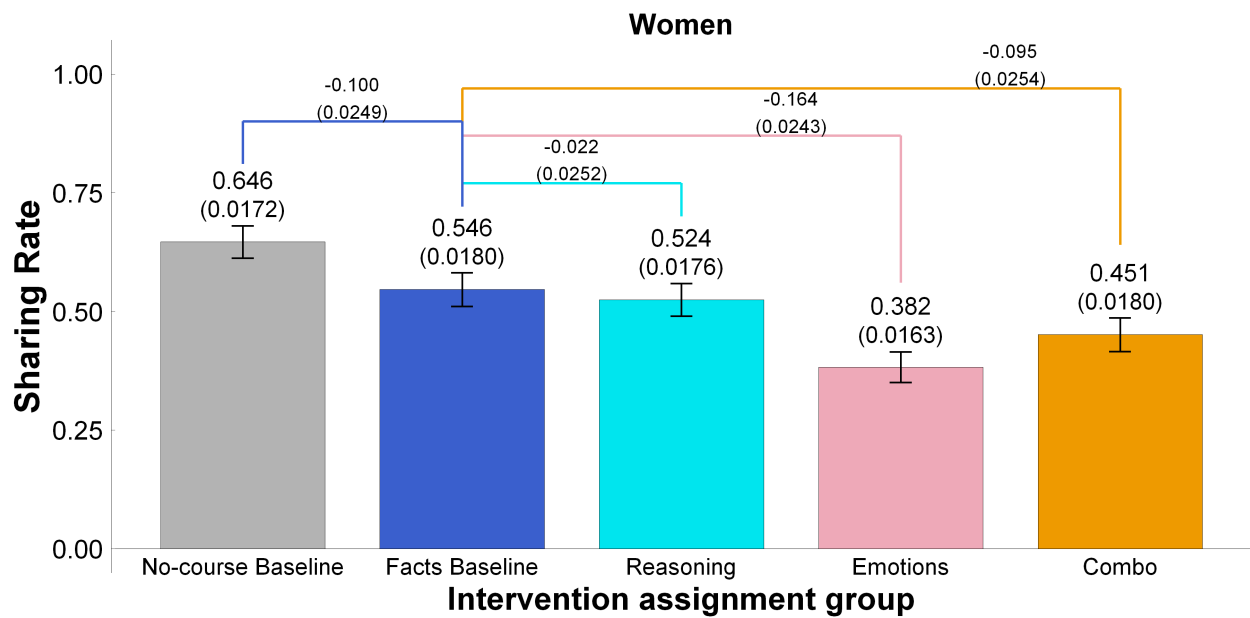


Figure E2: Primary outcome by treatment - Women

*Notes:* Sample includes the 3,111 participants who were assigned to the Reasoning or Emotions courses and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each bar displays the Sharing Rate for misinformation posts, as defined in Equation 1, by participants in their respective intervention assignment group. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	RW Adj p-Value
<b>Man Reasoning</b>	<b>Man Emotions</b>	977	1,038	0.481 (0.0127)	0.456 (0.0122)	0.025 (0.0177)	1 [0.1557]
<b>Woman Reasoning</b>	<b>Woman Emotions</b>	521	575	0.524 (0.0176)	0.382 (0.0163)	0.142 (0.0128)	0.0020 [0.0000]
<b>Difference-in-Differences</b>						-0.1169 (0.0298)	0.0290 [0.0001]
<b>Difference-in-Differences (percentage)</b>						-0.262 (0.0653)	0.0320 [0.0187]

Table E1: Subgroup analysis - Gender, Emotions vs Reasoning

*Notes:* Sample includes the 3,111 participants who were assigned to the Reasoning or Emotions courses and shared at least one non-misinformation post in the pre-survey, separately by gender and pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the results of t-tests. For the assignment groups specified in the first and second columns of the table, the t-tests compare the Sharing Rate for misinformation posts, as defined in Equation 1. Standard errors are in parentheses. Unadjusted p-values are in brackets. Romano-Wolf adjusted p-values are computed using 1,000 simulations and therefore have a minimum of 0.001. Such adjustment was performed separately for differences-in-means, differences-in-differences, and differences-in-differences (percentage) hypotheses, and including the covariates gender, age, prop of content shared, hours per day on social media, and pre-survey non-misinformation sharing rate. Full subgroup results available upon request.

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Mean 1	Mean 2	Diff. in Means	Unadj p-Value
<b>Gender – Followup Outcome</b>							
<b>Man Reasoning</b>	<b>Man Emotions</b>	552	563	0.427 (0.0174)	0.393 (0.0174)	0.034 (0.0246)	0.1701
<b>Woman Reasoning</b>	<b>Woman Emotions</b>	303	302	0.458 (0.0233)	0.368 (0.0233)	0.091 (0.0330)	0.0058
<b>Difference-in-Differences</b>						-0.057 (0.0412)	0.1634
<b>Gender – Diff in Disc. Score</b>							
<b>Man Reasoning</b>	<b>Man Emotions</b>	1,087	1,182	0.390 (0.0577)	0.578 (0.0550)	-0.188 (0.0797)	0.0185
<b>Woman Reasoning</b>	<b>Woman Emotions</b>	599	650	0.296 (0.0799)	0.685 (0.0771)	-0.389 (0.1110)	0.0005
<b>Difference-in-Differences</b>						-0.201 (0.1367)	0.1408
<b>Gender – Diff in Acc. Disc. Score</b>							
<b>Man Reasoning</b>	<b>Man Emotions</b>	1,087	1,182	1.323 (0.2585)	1.562 (0.2401)	-0.239 (0.3528)	0.4985
<b>Woman Reasoning</b>	<b>Woman Emotions</b>	599	650	0.838 (0.3602)	2.046 (0.3333)	-1.208 (0.4907)	0.0139
<b>Difference-in-Differences</b>						-0.969 (0.6044)	0.1089

Table E2: Subgroup analysis - Gender, Emotions vs Reasoning - Alternative outcomes

*Notes:* Follow-up sample includes the 1,720 participants who were assigned to the Reasoning or Emotions courses and shared at least one non-misinformation post in the post-survey, separately by gender and pooling participants in the Accuracy Inter and Accuracy After treatments. Discernment score and accuracy discernment score sample includes the 3,518 participants who completed the post-survey, separately by gender and pooling participants in the Accuracy Inter and Accuracy After treatments.

## E.2 Subgroup Analysis, by Post Type

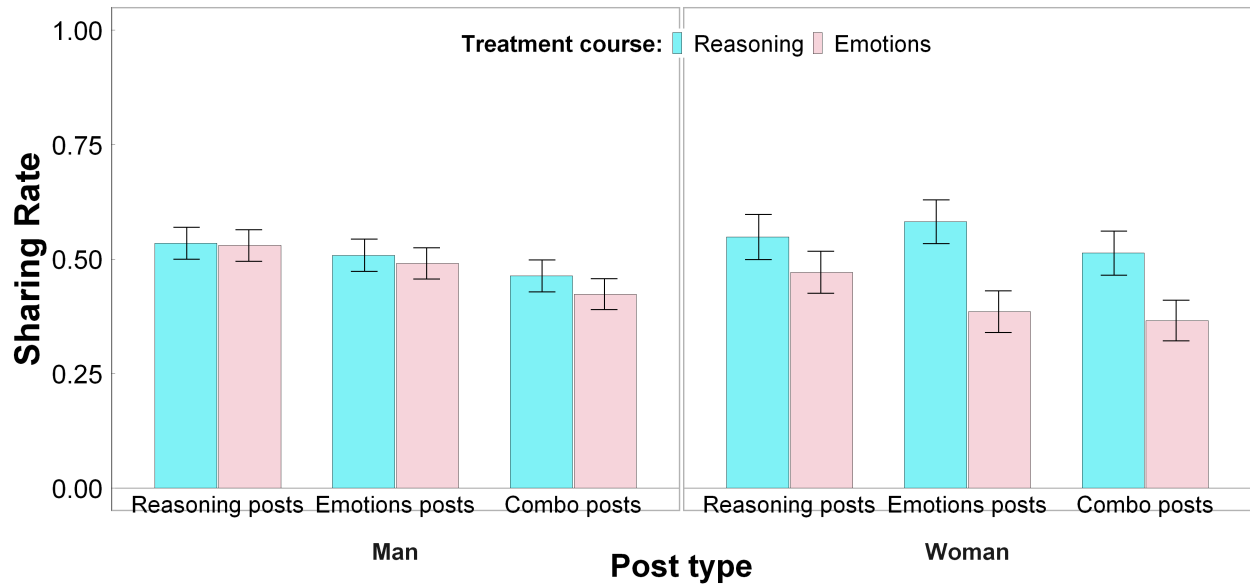


Figure E3: Misinformation Sharing, by Post Type and Gender

*Notes:* Sample includes 3,111 participants in the Reasoning or Emotions intervention assignment groups who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Panels are by gender, and each group of bars displays the Sharing Rate for misinformation posts of each type, as defined in Equation 1, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. Above each bar, the standard error is shown in parentheses below the Sharing Rate. The thin black bars represent 95% confidence intervals. Differences in Sharing Rates are shown above lines connecting the two relevant intervention assignment groups, with standard errors in parentheses.



### E.3 Subgroup Analysis, by Fact

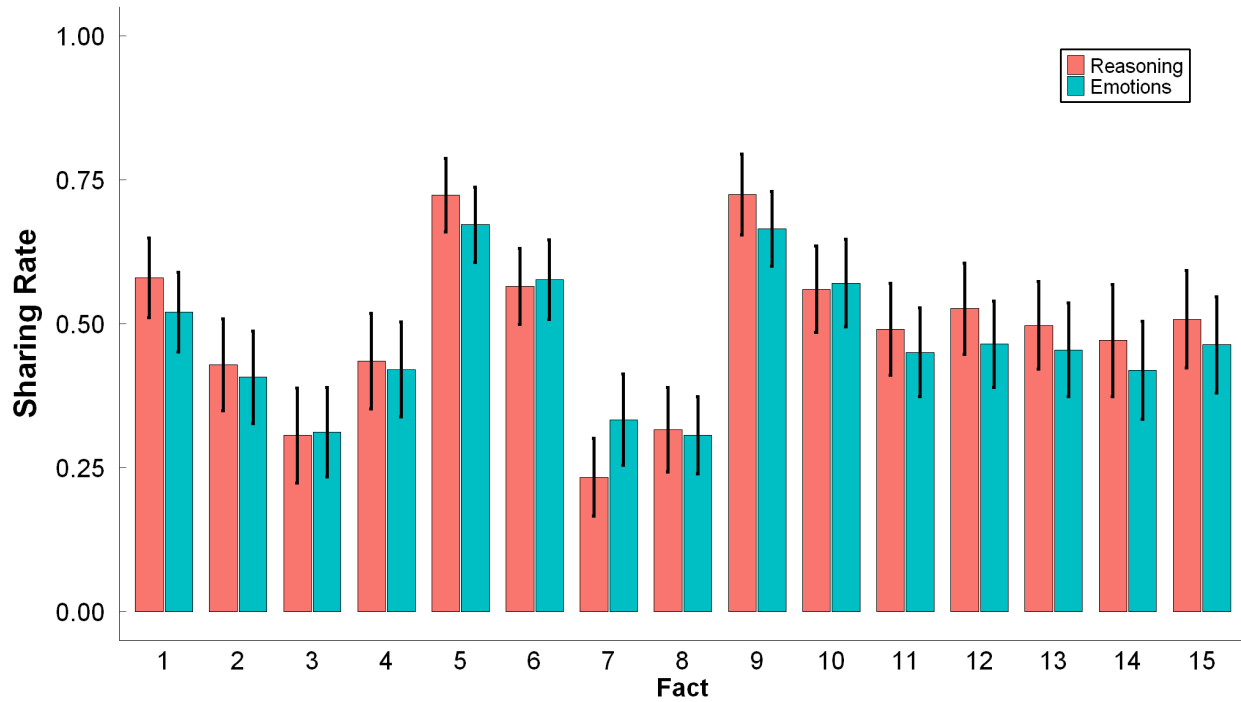


Figure E4: Misinformation Sharing, by Fact - Men

*Notes:* Sample includes 3,111 participants in the Reasoning or Emotions intervention assignment groups who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Each group of bars displays the Sharing Rate for misinformation posts of each fact, as defined in Equation 1, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. The thin black bars represent 95% confidence intervals.

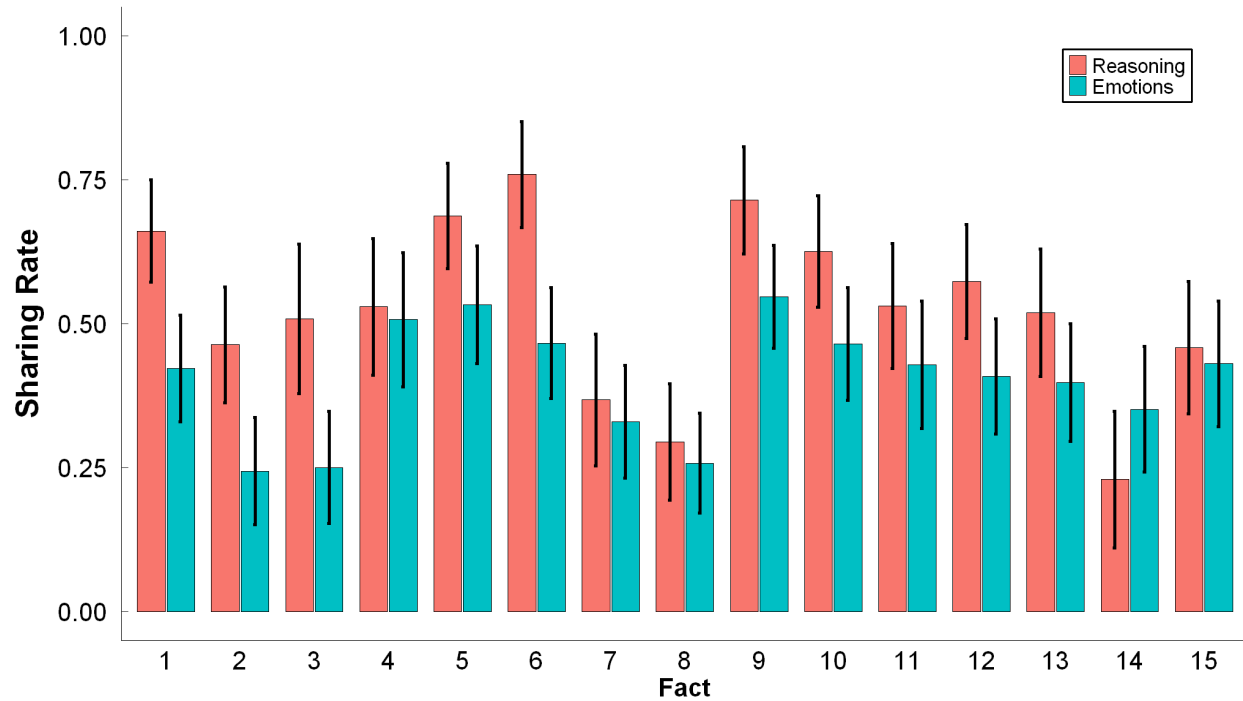


Figure E5: Misinformation Sharing, by Fact - Women

*Notes:* Sample includes 3,111 participants in the Reasoning or Emotions intervention assignment groups who completed the post-survey and shared at least one non-misinformation post in the pre-survey. Each group of bars displays the Sharing Rate for misinformation posts of each fact, as defined in Equation 1, by participants in their respective intervention assignment group, pooling participants in the Accuracy Inter and Accuracy After groups. The thin black bars represent 95% confidence intervals.

## E.4 Data-driven heterogeneity

Comparison	All covariates	Gender only
<b>Reasoning vs Emotions</b>	-0.011 (0.0211)	-0.036 (0.0115)
<b>Reasoning vs Combo</b>	0 (0)	0 (0)
<b>Emotions vs Combo</b>	-0.006 (0.0209)	-0.025 (0.0148)

Table E3: Rank-Weighted Average Treatment Effects on Sharing Rate

*Notes:* Sample includes the 7,688 participants who completed the post-survey and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row shows the estimate of the Rank-Weighted Average Treatment Effect (RATE) on the Sharing Rate for misinformation posts, as defined in 1, computed as the area under the Targeting Operator Characteristic curve as in [Yadlowsky et al. \(2021\)](#). The out-of-bag doubly robust scores and the out-of-sample Conditional Average Treatment Effects used as priorities were obtained using causal forests from [Athey et al. \(2019\)](#) including all available covariates (Column 1) and only the covariate Gender (Column 2)

## F Robustness to controlling for covariates using Augmented Inverse Probability Weighting (AIPW)

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Estimate	Holm Adj p-Value
<b>Treatment courses</b>	<b>No-course</b>	4,648	1,456	-0.199 (0.0405)	0.0000 [0.0000]
<b>Treatment courses</b>	<b>Facts</b>	4,648	1,584	-0.128 (0.0412)	0.0028 [0.0009]
<b>Reasoning</b>	<b>No-course</b>	1,502	1,456	-0.146 (0.0144)	0.0000 [0.0000]
<b>Emotions</b>	<b>No-course</b>	1,617	1,456	-0.211 (0.0140)	0.0000 [0.0000]
<b>Combo</b>	<b>No-course</b>	1,529	1,456	-0.184 (0.0145)	0.0000 [0.0000]
<b>Reasoning</b>	<b>Facts</b>	1,502	1,584	-0.073 (0.0146)	0.0000 [0.0000]
<b>Emotions</b>	<b>Facts</b>	1,617	1,584	-0.136 (0.0141)	0.0000 [0.0000]
<b>Combo</b>	<b>Facts</b>	1,529	1,584	-0.109 (0.0147)	0.0000 [0.0000]
<b>Facts</b>	<b>No-course</b>	1,584	1,456	-0.073 (0.0144)	0.0000 [0.0000]
<b>Reasoning</b>	<b>Emotions</b>	1,502	1,617	0.064 (0.0142)	0.0000 [0.0000]
<b>Reasoning</b>	<b>Combo</b>	1,502	1,529	0.036 (0.0147)	0.0275 [0.0138]
<b>Emotions</b>	<b>Combo</b>	1,617	1,529	-0.028 (0.0143)	0.0478 [0.0478]

Table F1: AIPW Average Treatment Effects: Misinformation Sharing

*Notes:* The sample is composed of the 7,688 participants who completed the post-survey and shared at least one non-misinformation post in the pre-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the Augmented Inverse Propensity Weighting (Robins et al., 1994) treatment effect estimates for the Sharing Rate for misinformation posts, as defined in 1. Both the propensity score and outcome predictions are obtained through probability forests (Athey et al., 2019). All predictions are out-of-sample predictions. Standard errors between parentheses. Unadjusted p-Values between brackets. Holm adjusted p-values are computed adjusting for all the tests presented in this table.

Group 1	Group 2	N. Obs. 1	N. Obs. 2	Estimate	Holm Adj p-Value
<b>Treatment courses</b>	<b>Facts</b>	2,554	857	-0.107 (0.0571)	0.0913 [0.0304]
<b>Reasoning</b>	<b>Facts</b>	857	857	-0.071 (0.0200)	0.0012 [0.0002]
<b>Emotions</b>	<b>Facts</b>	865	857	-0.127 (0.0199)	0.0000 [0.0000]
<b>Combo</b>	<b>Facts</b>	832	857	-0.078 (0.0202)	0.0004 [0.0001]
<b>Facts</b>	<b>No-course</b>	857	840	0.021 (0.0203)	0.5949 [0.2974]
<b>Reasoning</b>	<b>Emotions</b>	857	865	0.055 (0.0198)	0.0262 [0.0052]
<b>Reasoning</b>	<b>Combo</b>	857	832	0.009 (0.200)	0.6531 [0.6531]
<b>Emotions</b>	<b>Combo</b>	865	832	-0.046 (0.0199)	0.0801 [0.0200]

Table F2: AIPW Average Treatment Effects: Misinformation Sharing - Follow-up

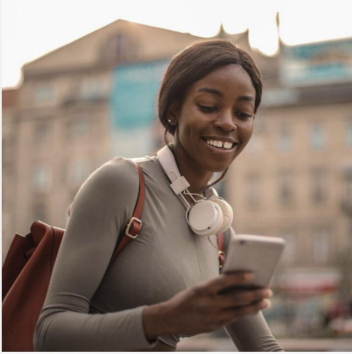
*Notes:* The sample is composed of the 4,251 participants who completed the follow-up survey and shared at least one non-misinformation post in the post-survey, pooling participants in the Accuracy Inter and Accuracy After treatments. Each row displays the Augmented Inverse Propensity Weighting (Robins et al., 1994) treatment effect estimates for the Sharing Rate for misinformation posts, as defined in 1, -but using instead the non-misinformation posts in the post-survey and their corresponding misinformation posts in the follow-up survey-. Both the propensity score and outcome predictions are obtained through probability forests (Athey et al., 2019). All predictions are out-of-sample predictions. Standard errors between parentheses. Unadjusted p-Values between brackets. Holm adjusted p-values are computed adjusting for all the tests in this table.

# Online Appendix A: Facebook Recruitment Ads Examples

**Social Impact Research Lab** Sponsored · 🌐 ...

Let's make social media better together!

Earn 🎧 AIRTIME 🎧 by signing up for our FREE, five-day text message course and completing short course surveys.



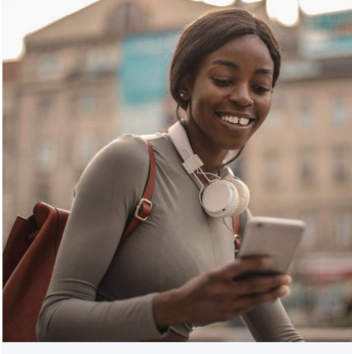
stanforduniversity.qualtrics.com  
**Start now** [Sign up](#)

👍 Joy Muthoni Muir... 11 Comments

👍 Like    💬 Comment    ➦ Share

**Social Impact Research Lab** Sponsored · 🌐 ...

Earn 🎧 AIRTIME 🎧 by signing up for our FREE, five-day text message course and completing short course surveys.



stanforduniversity.qualtrics.com  
**Start now** [Sign up](#)

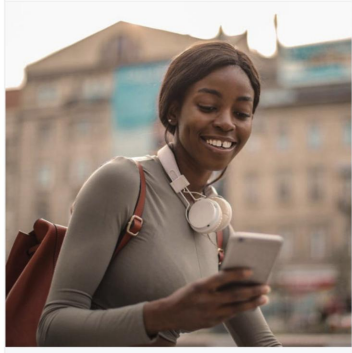
👍 Thomas Arisa and 2... 2 Comments 1 Share

👍 Like    💬 Comment    ➦ Share

**Social Impact Research Lab** Sponsored · 🌐 ...

Do you want to protect your family and friends from misinformation online?

Earn 🎧 AIRTIME 🎧 by signing up for our FREE, five-day text message course and completing short course surveys.



stanforduniversity.qualtrics.com  
**Start now** [Sign up](#)

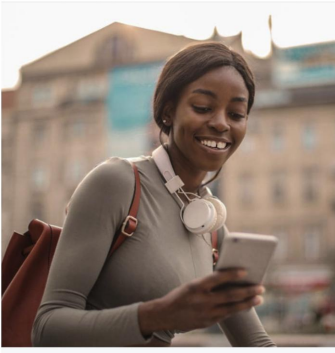
👍 Magdaline Adero and 9... 4 Comments

👍 Like    💬 Comment    ➦ Share

**Social Impact Research Lab** Sponsored · 🌐 ...

Not sure how to protect yourself from misinformation?

Earn 🎧 AIRTIME 🎧 by signing up for our FREE, five-day text message course and completing short course surveys.



stanforduniversity.qualtrics.com  
**Start now** [Sign up](#)

👍 Eli Muthaura and 25 others

👍 Like    💬 Comment    ➦ Share

# Online Appendix B: Text Message Courses

## Emotions Course

### DAY 1: Why people share misleading content

Let's start by talking about some definitions. Do you know the difference between disinformation and misinformation?

The difference is INTENT.

👉 DISINFORMATION is false or misleading information that is deliberately created and shared by people who know that it is false.

👉 MISINFORMATION is false or misleading information that people share without necessarily knowing that it is false.

Sometimes people share misleading information because they are trying to help, but it's important to remember the harm that can be caused whenever we share anything that's not 100% true.

? Let's test what you just learned. Pretend you have just seen a post on Facebook from your uncle sharing what he believes is an alternative treatment for Covid-19. You know that this treatment has been disproven. Is this an example of misinformation or disinformation?

A. Misinformation

✅ Correct! This is an example of MISINFORMATION. More likely than not, if your uncle shares an alternative treatment for Covid-19, he is probably just trying to help.

B. Disinformation

❌ Almost! This is an example of misinformation. More likely than not, if your uncle shares an alternative treatment for Covid-19, he is probably just trying to help. Disinformation is false or misleading information that is deliberately shared by people who know that it is false.

-----

There are a lot of reasons why people are trying to deliberately mislead you:

- 💰 To make money
- 🏆 To build reputation
- 😈 To cause trouble
- 🗳️ Political gain

➡️ Next time you see something misleading online, ask yourself if it's MISinformation or DISinformation, and think about why that person might have shared it!

Tomorrow, we'll talk about how our brains process misinformation. We'll also learn about how we can watch out for it.

Goodbye for now! 🙋

## **DAY 2: How to protect ourselves**

Let's talk about our brains. 🧠

🧑 Our brains help us make decisions all day long. But there are also some things they do that make us more susceptible to misinformation.

### **1) EMOTIONAL TRIGGERS**

— Ever heard the phrase "fight or flight"? When we're scared, we do whatever we need to protect ourselves and others. 🙋 That includes consuming and sharing information we believe will keep us safe but which might not actually be true.

— The same is true for information that makes us angry, 😡 scared 😱 or sad 😞. When we have strong emotional reactions, we're less likely to stop, think and analyze what we're seeing.

### **2) MENTAL SHORTCUTS**

— We also use a lot of mental shortcuts to help us make snap decisions. 🏃 For example, if a friend posts something misleading on social media, we might be more likely to share it because we trust that person.



— Here's the problem: People who spread disinformation often rely on these emotional triggers and mental shortcuts to manipulate us. 😞 Your friend might not know they are being misled, but you can help stop the spread by not sharing. 🛑

? How about you? If a family member or a friend posts something online, do you stop to check whether it's misinformation?

A. I always trust my family and friends!

That's normal! But remember: Everyone makes mistakes. It's always a good idea to verify whether something is true before resharing it. ✅

B. I always check whether something is misinformation.

That's great! Keep it up. 🎯

➡ One of the best ways to protect ourselves from misinformation is to recognize when we're having a strong emotional response. Try following these steps:

1 🛑 STOP: If you read or see something online that makes you feel strongly, pause before sharing it.

2 😞 QUESTION: Is the post's creator trying to manipulate how you think about something? Is the information misleading? If there is any doubt in your mind, do NOT share it.

Remember that by sharing misleading information, you could cause real harm. Just like you trust many of your friends and family, they trust you too! 💪

See you tomorrow! 🙌

### DAY 3: Fear



It's time for our next lesson! Over the next three days, we're going to show you a few examples of how misinformation targets your emotions.

Today let's talk about FEAR.

🧠 Life can seem pretty scary right. Feeling scared or uncertain is rational, but it's important to understand that when we are scared we are also more vulnerable to being manipulated. 🚫

Take a look at the image above.

Remember the two steps?

1 🛑 STOP: The image suggests that the ingredients in the medicine are unnatural or unsafe — that sounds scary! But let's take a second to think about it.

2 😟 QUESTION: Is the post's creator trying to manipulate how you think about the medicine?

? What motive might someone have for wanting you to believe the ingredients of another product are safer?

A. To sell their product.

- B. Because they genuinely think the other product is dangerous.
- C. All of the above.

The correct answer is all of the above! ✅ In this example, the person who made this image is clearly trying to make a profit off the alternative product. 📊 But, of course, the person also might genuinely think the other product is dangerous.

➡ In either case, notice how the image makes you FEEL. Health information can be difficult to understand because of unfamiliar terms, and unfamiliarity often breeds fear. 🧑

But remember: Fear makes you more vulnerable to being manipulated, and people who spread disinformation often use this to their advantage.

Stop and seek out all the information you might need to make informed decisions. And if in doubt, do NOT share. 🙅

That's all for now! See you tomorrow. 🙌

#### Day 4: Anger

**DON'T LET THE GOVERNMENT MAKE  
YOUR CHILD AN EXPERIMENT**



Welcome back! Today let's talk about ANGER.

Our beliefs are part of our identity, and when these beliefs are attacked we often become angry and defensive. 😡

👉 Anger is also a mobilizing force. Research shows that misinformation travels much faster on social media if it provokes anger.

Some platforms have even prioritized showing users content that elicits angry reactions! This makes it much harder for us to avoid content that is polarizing. 🙄

? Take a look at the image above. Would you share this on your social media feed?

- A. Yes.
- B. No.

Before sharing, let's stop and question all the ways this post tries to make us feel anger:

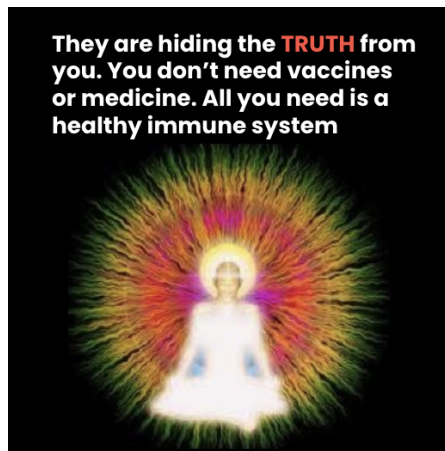
- ➡ The picture shows a crying child. People are naturally protective of their children.
- ➡ The caption implies that the vaccination being administered is experimental.
- ➡ The caption also implies that the "government" does not care about children's health or safety.

Instead of providing useful information to help you understand whether the vaccine depicted is safe and why, this post is relying on emotional triggers to make you angry.

🗨️ Remember: Before sharing something that makes you angry, STOP and QUESTION the information. Don't share anything if you're not 100% sure it is true.

See you for our final lesson tomorrow! 🙌

## Day 5: Superiority



Hi again! Today is our last day, and we're going to talk about SUPERIORITY.

There may come a time when you believe that you are privy to information that no one else has access to — that you are part of a select few who sees the “truth.” 👁

It's important to notice when this happens! Here's why:

🧠 Our brains look for clues to quickly work out whether something is credible.

👉 Having access to insider or expert knowledge about a particular issue, or the illusion of having access to this information, is one example of this.

These feelings of “knowing more” than other people might make us feel good, but they can also make us less critical and more prone to believing misinformation. ⚠

➡ Here's one tip. Watch out for health information on social media that begins with language like this:

—1 “My friend is a nurse and she said...”

—2 “My brother works for the government and has inside knowledge. He just told me that...”

🤔 The next time you're presented with "insider information" that someone claims is from an expert, try to understand WHY that person is an expert and see whether there are other people in their field who agree with them.

-

! IMPORTANT ! In order to complete the course and receive free airtime, take a second to fill out our final survey. Are you ready?

A. Yes

Great! Here is the link:

B. No

That's OK! We will remind you again tomorrow. Here's the link if you change your mind.

#### **DAY 6: End-of-course survey**

! IMPORTANT ! Did you fill out the end-of-course survey yesterday? In order to receive KSH 500 in AIRTIME, you MUST fill out the survey:

Completing the survey also helps us improve our course and research. 🙏

Congratulations! In this course, you've learned about how misinformation spreads, how it can trick your brain and how to avoid it. 🎉

The next time you're scrolling through your social media feeds and something jumps out at you, stop, think and remember some of the tips you've learned. Doing so could protect you and the ones you love from possible manipulation. 💪

Goodbye! 👐

# Info Course

## DAY 1: Why people share misleading content

Let's start by talking about some definitions. Do you know the difference between disinformation and misinformation?

The difference is INTENT.

👉 DISINFORMATION is false or misleading information that is deliberately created and shared by people who know that it is false.

👉 MISINFORMATION is false or misleading information that people share without necessarily knowing that it is false.

Sometimes people share misleading information because they are trying to help, but it's important to remember the harm that can be caused whenever we share anything that's not 100% true.

? Let's test what you just learned. Pretend you have just seen a post on Facebook from your uncle sharing what he believes is an alternative treatment for Covid-19. You know that this treatment has been disproven. Is this an example of misinformation or disinformation?

A. Misinformation

✅ Correct! This is an example of MISINFORMATION. More likely than not, if your uncle shares an alternative treatment for Covid-19, he is probably just trying to help.

B. Disinformation

❌ Almost! This is an example of misinformation. More likely than not, if your uncle shares an alternative treatment for Covid-19, he is probably just trying to help. Disinformation is false or misleading information that is deliberately shared by people who know that it is false.

There are a lot of reasons why people are trying to deliberately mislead you:

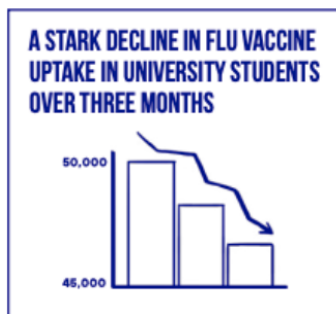
- 💰 To make money
- 👤 To build reputation
- 😡 To cause trouble
- 📊 Political gain

➡ Next time you see something misleading online, ask yourself if it's MISinformation or DISinformation, and think about why that person might have shared it!

☀ Tomorrow, we'll talk a little bit about how our brains process misinformation. We'll also learn about how we can watch out for it.

Goodbye for now! 🙌

## DAY 2: Misleading graphs



🙌 Welcome back! Over the next few days, we're going to talk about clues and signs to watch for, so that you can get better at spotting mis- and disinformation. 🧐

As the old saying goes: "A picture is worth a thousand words." 🖼

⚠ This is especially true when it comes to misinformation.

📊 Graphs and diagrams often make information look more official. The danger is that they can easily be misleading, especially if they use cherry-picked statistics and don't provide all of the data and context.



Misleading graphs or diagrams may be spread intentionally. But even accurate diagrams, when done poorly, can be misinterpreted. 🧑

When you see a graph or diagram on social media, be sure to look it over carefully.

👉 Check the source, where the data comes from, as well as the context in which it was originally shared.

? Look at the chart above. Can you tell what's misleading about it?

- A. There has been no decline in flu vaccine uptake among university students
- B. The numbers have been fabricated
- C. The chart is set up in a way that exaggerates the decline

Automated response:

The correct answer is C! Look carefully at the vertical axis. It starts at 45,000 instead of 0. If the axis started at zero, the decline in vaccine uptake would look less dramatic! This often happens with infographics. The data being displayed might be correct, but the *way* it is displayed is misleading.

That's all for now! We'll see you tomorrow, with a new tactic to learn about. 🧠👉

### DAY 3: Websites and impostors

🖥️ Let's talk about ways people can make misleading content look more "official." A WEBSITE is one way to do this.

Websites are easy to make and they provide a quick way to monetize misinformation by connecting content to advertising. 🤖

📰 Websites that post misinformation are often made to look like an established news or health authority site. This is what we call "impostor content." 🤡

➡️ Impostor content is when somebody makes it look as if their information came from an organization people recognize and trust. Somebody could do this by using another brand's logo or simply having a similar name or visuals.

🔍 When you see an unfamiliar website or social media page that is pushing health information, try to figure out who created it.

- 🤔 What does the "About" page say?
- 🤔 Who contributes the information?
- 🤔 What are their credentials and background?

Here are some red flags:

- 🚩 The website uses content that comes from somewhere else
- 🚩 The website is selling a health product
- 🚩 The website URL has an unusual ending
- 🚩 The contributors claim to be doctors but you can't find their medical credentials

Goodbye for now! 🙋

#### **Day 4: Eyewitness media**

🙋 Welcome back!

Today we're going to talk about the power of eyewitness media. 👁

👉 Videos and audio that make it look as if someone experienced something first-hand can be VERY convincing.

👉 Throughout the Covid-19 pandemic, people have shared videos they claim to have filmed showing empty hospitals or people having adverse reactions to vaccines.

These videos are often highly emotive and may depict shocking material. ⚠️ But they are also frequently taken out of context! Here are two examples:

- ➡️ The video or picture might be real, but the caption is misleading.
- ➡️ The video or picture might be old, and it's being reshared to represent something happening now. 🕒

It's always a good idea to verify pictures and videos before you reshare them. Here are some questions to ask yourself:

- ✓ Is the person sharing a picture or video the person who took the picture or video?
- ✓ Is it possible that the picture or video came from somewhere else?
- ✓ If so, where did it come from? Can you find other examples of it on the internet?
- ✓ When was the picture or video taken?

--

? Want to know about a tool you can use to look up where a photo came from?

A. Yes! Show me the tool.

This is a free tool that works on both Google Chrome and Firefox. When you install it, it allows you to right-click on any photo and do what's called a reverse image search.

<https://chrome.google.com/webstore/detail/reveye-reverse-image-sear/keaacjcjhehbbapnp/hnmpiklalfhelgf?hl=en>

B. That's OK! I've learned enough for one day.

OK, great!

Bye for now! 🙌

### Day 5: The kernel of truth

#### CONCERT IN NAIROBI

Are they crazy? No one in this photo has a mask on!



Today we're going to talk about one of the most important concepts: CONTEXT.

🤖 The reality is that most misinformation isn't entirely false. A lot of misinformation has a small kernel of truth that has been presented in a false or misleading way. 🔍

!! 🗣️ This is important, because it's often the small kernel of truth that makes the content more believable.

💬 Examples of this kind of misinformation include content taken out of context, such as a quote that is cut and reframed to change the meaning. ✂️

Because the content contains some degree of truth, it is more complicated to debunk.

➡️ Remember: Context is important. Ask yourself not only whether particular facts are true, but also whether they are being presented accurately.

—

? Say you come across an image like the one above. What do you think is the most likely explanation for why it's misleading?

- A. It has been Photoshopped
- B. It is out of context

Sometimes images are Photoshopped, but most of the misinformation you will come across is not outright fabricated, it's just out of context. This could be an old photo of a maskless crowd from 2015, depicted as if it had been taken during the pandemic.

! IMPORTANT ! In order to receive airtime, take a second to fill out our final survey. Type 'A' if you're ready for the link.

Great! Please take our end-of-course survey here:

#### **DAY 6: End-of-course survey**

! IMPORTANT ! Did you fill out the end-of-course survey yesterday? In order to receive airtime, you MUST fill out the survey.

🙏 Completing the survey also helps us improve our course and research.

Congratulations! In this course, you've learned about how misinformation spreads, how it can trick your brain and how to avoid it. 🎉

The next time you're scrolling through your social media feeds and something jumps out at you, stop, think and remember some of the tips you've learned. Doing so could protect you and the ones you love from possible manipulation.

Goodbye! 🙌

## Combo Course

### **DAY 1: Why people share misleading content**

Let's start by talking about some definitions. Do you know the difference between disinformation and misinformation?

The difference is INTENT.

👉 **DISINFORMATION** is false or misleading information that is deliberately created and shared by people who know that it is false.

👉 **MISINFORMATION** is false or misleading information that people share without necessarily knowing that it is false.

Sometimes people share misleading information because they are trying to help, but it's important to remember the harm that can be caused whenever we share anything that's not 100% true.

❓ Let's test what you just learned. Pretend you have just seen a post on Facebook from your uncle sharing what he believes is an alternative treatment for Covid-19. You know that this treatment has been disproven. Is this an example of misinformation or disinformation?

A. Misinformation

✅ Correct! This is an example of MISINFORMATION. More likely than not, if your uncle shares an alternative treatment for Covid-19, he is probably just trying to help.

#### B. Disinformation

❌ Almost! This is an example of misinformation. More likely than not, if your uncle shares an alternative treatment for Covid-19, he is probably just trying to help. Disinformation is false or misleading information that is deliberately shared by people who know that it is false.

There are a lot of reasons why people are trying to deliberately mislead you:

- 💰 To make money
- 👤 To build reputation
- 😡 To cause trouble
- 🗳️ Political gain

➡️ Next time you see something misleading online, ask yourself if it's MISinformation or DISinformation, and think about why that person might have shared it!

☀️ Tomorrow, we'll talk about how our brains process misinformation. We'll also learn about how we can watch out for misinformation.

👋 Goodbye for now!

### **DAY 2: How to protect ourselves**

Let's talk about our brains. 🧠

🧑 Our brains help us make decisions all day long. But there are also some things they do that make us more susceptible to misinformation.

#### 1 EMOTIONAL TRIGGERS

— Ever heard the phrase "fight or flight"? When we're scared, we do whatever we need to protect ourselves and others. 🧡 That includes consuming and sharing information we believe will keep us safe but which might not actually be true.

— The same is true for information that makes us angry, 😡 scared 😱 or sad 😞. When we have strong emotional reactions, we're less likely to stop, think and analyze what we're seeing.

## 2 MENTAL SHORTCUTS

— We also use a lot of mental shortcuts to help us make snap decisions. 🏃  
For example, if a friend posts something misleading on social media, we might be more likely to share it because we trust that person.

— Here's the problem: People who spread disinformation often rely on these emotional triggers and mental shortcuts to manipulate us. 😞 Your friend might not know they are being misled, but you can help stop the spread by not sharing. 🛑

? How about you? If a family member or friend posts something online, do you stop to check whether it's misinformation?

A. I always trust my family and friends!

That's normal! But remember: Everyone makes mistakes. It's always a good idea to verify whether something is true before resharing it. ✅

B. I always check whether something is misinformation.

That's great! Keep it up. 🎯

➡ One of the best ways to protect ourselves from misinformation is to recognize when we're having a strong emotional response. Try following these steps:

1 🛑 STOP: If you read or see something online that makes you feel strongly, pause before sharing it.

2 😞 QUESTION: Is the post's creator trying to manipulate how you think about something? Is the information misleading? If there is any doubt in your mind, do NOT share it.

Remember that by sharing misleading information, you could cause real harm. Just like you trust many of your friends and family, they trust you too! 🤝

### DAY 3: Paying attention to our emotions



Let's talk about some ways that misinformation targets your emotions.

- Fear
- Anger
- Superiority

#### 1 FEAR 🗿

Life can seem scary right now. Many people have been seriously ill in the pandemic, and we're not sure what the future holds.

🗿 Feeling scared or uncertain is rational, but it's important to understand that when we are scared we are also more vulnerable to being manipulated.

#### 2 ANGER 😡

It has also been shown that misinformation spreads faster if it provokes anger.

Some platforms have even prioritized showing users content that receives angry reactions! This makes it much harder for us to avoid content that is polarizing. 🗿



### ③ SUPERIORITY 😊

There may come a time when you believe that you are privy to information that no one else has access to — that you are one of a select few who see the “truth.” 👁

These feelings of “knowing more” might make us feel good, but they can also make us less critical ⚠

➡ Here’s one tip. Watch out for health information on social media that begins with language like this:

- ① “My friend is a nurse and she said...”
- ② “My brother works for the government. He told me...”

? Take a look at the image above. What motive might someone have for making you believe the ingredients of another product are safer?

- A. To sell their product.
- B. To make you scared of the other product.
- C. Because they genuinely think the other product is dangerous.
- D. All of the above.

The correct answer is all of the above! ✔ In this example, the person who made this image is clearly trying to make a profit off their alternative product. 🟩 But, of course, the person also might genuinely think the other product is dangerous!

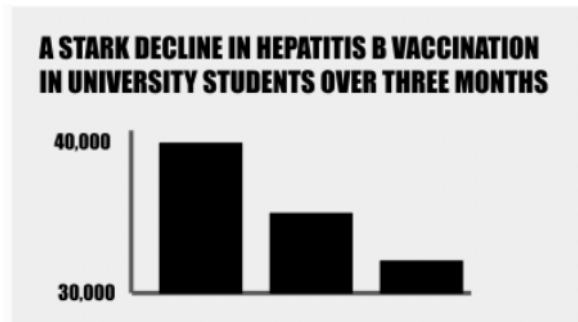
➡ In either case, notice how the image makes you FEEL. Health information can be difficult to understand because of unfamiliar terms, and unfamiliarity often breeds fear. 🧑

But remember: Fear makes you more vulnerable to being manipulated, and people who spread disinformation often use this to their advantage.

Stop and seek out all the information you might need to make informed decisions. And if in doubt, do NOT share. 🙅

Bye for now! 🙋

#### Day 4: Misinformation tactics



Let's talk about some of the things that can make misinformation look so credible.

- Websites and impostors
- Eyewitness media
- Graphs and diagrams

#### 1 WEBSITES AND IMPOSTORS

Websites are easy to make and they provide a quick way to monetize misinformation by connecting content to advertising. 🤖

📰 Websites that post misinformation are often made to look like an established news or health authority site. This is what we call "impostor content." 😡

#### 2 "EYEWITNESS" MEDIA

Videos that make it look as if someone experienced something first-hand can be VERY convincing. 📺 Throughout the Covid-19 pandemic, people have shared videos they claim to have filmed showing empty hospitals or people having adverse reactions to vaccines.

These videos often depict shocking material. ⚠️ But they are also frequently taken out of context!

#### 3 GRAPHS AND DIAGRAMS

📊 Graphs and diagrams often make information look more official. The danger is that they can easily be misleading, especially if they use cherry-picked statistics and don't provide all of the data and context.

Even accurate diagrams, when done poorly, can be misinterpreted. 🤖

- ? Take a look at the chart above. Can you tell what's misleading about it?
- A. There has been no decline in flu vaccine uptake among university students
  - B. The numbers have been fabricated
  - C. The chart is set up in a way that exaggerates the decline

Automated response:

The correct answer is C! Look carefully at the vertical axis. It starts at 45,000 instead of 0. If the axis started at zero, the decline in vaccine uptake would look less dramatic! This often happens with infographics. The data being displayed might be correct, but the *way* it is displayed is misleading.

That's all for now! See you tomorrow for our last lesson. 🍌

### Day 5: The kernel of truth

#### CONCERT IN NAIROBI

Are they crazy? No one in this photo has a mask on!



Today we're going to talk about one of the most important concepts: CONTEXT.

🤖 The reality is that most misinformation isn't entirely false. A lot of misinformation has a small kernel of truth that has been presented in a false or misleading way. 🔍

!! 🗣️ This is important, because often the small kernel of truth makes the content more believable.

💬 Examples of this kind of misinformation include content taken out of context, such as a quote that is cut and reframed to change the meaning. ✂️

Because the content contains some degree of truth, it is more complicated to debunk.

➡️ Remember: Context is important. Ask yourself not just whether particular facts are true, but also whether they are being presented accurately.

–

? Say you come across an image like the one above. What do you think is the most likely explanation for why it's misleading?

- A. It has been Photoshopped
- B. It is out of context

Sometimes images are Photoshopped, but most of the misinformation you will come across is not outright fabricated, it's just out of context. This could be an old photo of a maskless crowd from 2015 depicted as if it had been taken during the pandemic.

! IMPORTANT ! In order to receive airtime for taking this course, take a second to fill out our final survey.

#### **DAY 6: End-of-course survey**

! IMPORTANT ! Did you fill out the end-of-course survey yesterday? In order to receive KSH 500 in AIRTIME, please take a second to fill out our final survey:

🙏 Completing the survey also helps us improve our course and research.

Congratulations! In this course, you've learned about how misinformation spreads, how it can trick your brain and how to avoid it. 🎉

The next time you're scrolling through your social media feeds and something jumps out at you, stop, think and remember some of the tips you've learned. Doing so could protect you and the ones you love from possible manipulation.

Goodbye! 🙌

## Facts Baseline Course

### DAY 1:

? Did you know that "lies spread faster than the truth" on social media?

📄 A study from 2018 examined how both true and false news stories spread on Twitter. The authors of the study showed how false news stories traveled farther and faster on the social media platform than verified news stories.

📄 They also showed that of all the different kinds of false news stories, such as false news stories about science or natural disasters, false political news spread the fastest.

📄 Did you also know that being proficient with social media does not necessarily correlate to being skillful in detecting misinformation?

📄 A study from 2021, showed that high school students in the US, despite their frequent use of social media, were largely unable to discern whether a piece of content was true or false.

-

? Let's test what you just learned. Does being proficient on social media correlate to being good at detecting misinformation?

A. Yes

✗ It turns out there is NOT a correlation between being proficient on social media and being skillful in detecting misinformation. A study 📄 from 2021, showed that high school students in the US, despite their frequent use of social media, were largely unable to discern whether a piece of content was true or false.

B. No, not necessarily

✅ That's right. There is NOT a correlation between being proficient on social media and being skillful in detecting misinformation. A study 📄 from 2021, showed that high school students in the US, despite their frequent use of social media, were largely unable to discern whether a piece of content was true or false.

## DAY 2:

Many people may believe that misinformation and disinformation are new phenomena that came about with the rise of social media.

!! But misinformation isn't exactly a new thing.

👤 Social media and advanced communication technology may have allowed for misinformation to circulate wider and faster than ever before, but misinformation and disinformation have a long history alongside humankind.

👤 During the Roman Empire, for example, leaders used misinformation to come to power. Nazi Germany was known for its use of propaganda and disinformation. And much of the sensationalist tabloid or "yellow journalism", which effectively treated rumors and falsehoods as facts, was popular in the mid-1900s.

-

Is misinformation a new phenomena?

A. Yes

❌ Misinformation has actually been around for a long time. Yet, social media and advanced communication technology 📱 may have allowed for misinformation to circulate wider and faster than ever before.

B. No, not necessarily

✅ That's right. Misinformation has a long history alongside humankind. Leaders of the Roman Empire used it as did Nazi Germany. There are many historical examples of misinformation.

## DAY 3:

😬 For those of you who think you are unaffected by false information or would know if misinformation was influencing your behavior, guess again!!

📖 Several studies have shown that disinformation, even the slightest amount, can affect your unconscious behavior in very subtle ways.

🔥 Misinformation can also influence your reasoning and decision-making resulting in poor judgment.

🧠 Because of the way our brain operates, even when misinformation we have been exposed to is corrected with factual information, we often continue to be biased towards misinformation. It's what psychologists refer to as the "continued influence effect".

-

? Pretend someone was exposed to a piece of misinformation which they believed. The next day they read a fact-check that debunks the misinformation. What happens to that person?

A. The person is no longer influenced by the piece of misinformation.

✗ Because of the way our brain 🧠 operates, even when misinformation we have been exposed to is corrected with factual information, we often continue to be biased towards misinformation. It's what psychologists refer to as the "continued influence effect".

B. The person continues to be influenced by misinformation.

✓ That's right. Because of the way our brain 🧠 operates, even when misinformation we have been exposed to is corrected with factual information, we often continue to be biased toward misinformation. It's what psychologists refer to as the "continued influence effect".


#### DAY 4:


👤 In the last lesson, we learned about the "continued influence effect", which shows how we continue believing misinformation even after it has been corrected.

😊 In today's lesson we will learn about another phenomenon called the "third-person effect", which is particularly relevant to all of us who think that we are immune to misinformation.




👉 The "third-person effect" describes how individuals perceive the effects of misinformation to be much more powerful on others than on themselves. In essence, we think that everyone is susceptible to misinformation except ourselves.


📖 Yet, studies show that people who are the most confident at being able to spot misinformation are the ones most likely to believe it. This is especially true when we are subjected to information overload — when we feel overwhelmed by the amount of information we are exposed to.

 Keep that in mind the next time you are reading the news on social media or have hundreds of tabs open on your computer.




-  
 How accurate are humans when it comes to determining the effects of misinformation on ourselves?

A. Super accurate.


 You are super wrong. The “third-person effect” describes how individuals perceive the effects of misinformation to be much more powerful on others  than on themselves . In essence, we think that everyone is susceptible to misinformation except ourselves.

We think that everyone is susceptible to misinformation except ourselves.  
Yet, studies  show that people who are the most confident at being able to spot misinformation are the ones most likely to believe it.




B. Pretty accurate.

 Not so much. The “third-person effect” describes how individuals perceive the effects of misinformation to be much more powerful on others  than on themselves . In essence, we think that everyone is susceptible to misinformation except ourselves.


We think that everyone is susceptible to misinformation except ourselves.

Yet, studies  show that people who are the most confident at being able to spot misinformation are the ones most likely to believe it.


C. Not so accurate.

 That's correct. The “third-person effect” describes how individuals perceive the effects of misinformation to be much more powerful on others  than on themselves . In essence, we think that everyone is susceptible to misinformation except ourselves.


We think that everyone is susceptible to misinformation except ourselves.

Yet, studies  show that people who are the most confident at being able to spot misinformation are the ones most likely to believe it.


#### DAY 5:


 We may think of misinformation as being specific to a particular context or country.




 For example, you might think that the specific rumors and misinformation about a particular medicine or vaccine in one country wouldn't appear in another country.

But the internet and social media are, to a large degree, borderless. Information, including misinformation and disinformation, travels seamlessly between countries and continents.

 Language is one good predictor of where misinformation may spread to. Misinformation in one language, such as English, will more easily make its way to other countries and communities that speak or understand English.

 But language isn't the only predictor. If a piece of misinformation is popular enough, it is often translated word-for-word into other languages, taking on a life of its own.

 Disinformation campaigns are no different. Where bad actors in different countries share similar beliefs or ideologies, they may help each other to influence matters in those countries.

Misinformation is truly a global phenomenon affecting all of us.

-

**! IMPORTANT !** In order to complete the course and receive KSH 500 in AIRTIME, take a second to fill out our final survey. Are you ready?

A. Yes.


Great! Here is the link:

B. Not yet.

That's okay! We will remind you again tomorrow. Here's the link if you change your mind.

## **DAY 6:**

**! IMPORTANT !** Did you fill out the end-of-course survey yesterday? In order to receive KSH 500 in AIRTIME, please take a second to fill out our final survey.

 Congratulations! In this course, you've learned about how misinformation spreads, how it can trick your brain, and how to avoid it.

The next time you're scrolling through your social media feeds and something jumps out at you, stop, think and remember some of the tips you've learned. Doing so could protect you and the ones you love from possible manipulation.

🙏 If you haven't already, please help us improve our course and research by completing our end-of-course survey:

🎁 If you would like to receive KSH 500 in AIRTIME for completing this course, you **MUST** fill out the survey.

# Online Appendix C: Pre-survey Screenshots

## F.1 Main Survey

WELCOME!

You are invited to participate in a research study on news. There will be two surveys and a short and free text message course. Your participation will take approximately 10 - 15 minutes for each survey. You will receive the second survey after approximately 5 days after you finish this survey.

You will receive **KSH 500 in AIRTIME** as a thank you for your participation in this research study. You'll be paid within 24 hours of completing the second survey.

**Also, make sure you have not completed this survey before. If you complete it more than once, you will not be able to complete the study and receive AIRTIME.**

Before we start, please read the consent form and confirm you want to participate in the next screen.



## CONSENT FORM

This is a minimal-risk study, and all of your information is confidential. We cannot promise that you will receive any benefits from this study. Your participation is voluntary and you have the right to refuse to participate, skip or refuse to answer any particular questions, or stop at any time.

If you have any questions or concerns, you can contact the Protocol Director, Herman Donner, at [inoculationagainstmisinfo@gmail.com](mailto:inoculationagainstmisinfo@gmail.com) / [hdonner@stanford.edu](mailto:hdonner@stanford.edu) or our research review board to speak to someone independent of the research team at [irbnonmed@stanford.edu](mailto:irbnonmed@stanford.edu).

---

If you agree to participate in this research and confirm you are 18 years or older, please click "Yes, I agree".

Yes, I agree

No



You will see 10 social media posts on the following screens. Please read each one of them and answer the question at the end of each screen.

---

We will include **AT LEAST ONE attention check**, which will be in the format of a social media post. When you see the attention check(s), choose the answer as instructed in the post.

**If you don't answer ALL the attention checks correctly, you will not receive AIRTIME regardless of whether you completed the survey.**

---

**Also, make sure you have not completed this survey before. If you complete it more than once, you will not be able to complete the study and receive AIRTIME.**

---

Please confirm your understanding of the instructions above.

- I acknowledge that I will not receive airtime if I do not pass the attention checks, regardless of whether I complete the survey or not, or if I complete this survey more than once.



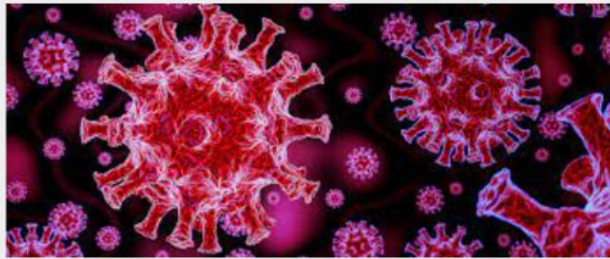
Please read the post and answer the question below.



Ali Ouma  
8 hrs



This is OUTRAGEOUS! The government is pushing COVID-19 vaccines that don't even protect against deadly COVID-19 infections.



To the best of your knowledge, how accurate is the claim in the above post?

Not at all accurate

Not very accurate

Somewhat accurate

Very accurate



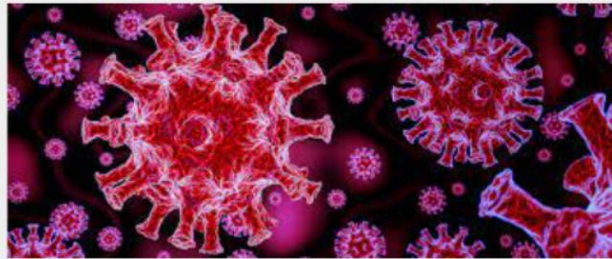
Please read the post and answer the question below.



**Ali Ouma**  
8 hrs



This is **OUTRAGEOUS!** The government is pushing COVID-19 vaccines that don't even protect against deadly COVID-19 infections.



Would you share this post?

No

Yes





Grace Wangari

5 hrs



Lupita Nyong'o looks gorgeous for 39 years old! Apparently her diet (no sugar, limited carbs and calories, low inflammatory, etc) is the latest in longevity science. Inspiring!



In a previous question, you said that you did not want to share this post. What made you NOT want to share it?







**Samwel Wambui**

10 hrs



I just read on Nation that fewer than 1% in the world have [#autism](#), according to the WHO. In Kenya, that rate is 4%. Scientists say that they can't still find the exact causes of [#autism](#), but I don't buy any of that, and neither should you. Who is responsible for this? Share to raise awareness!

[The reality of living with autism in Kenya | Nation](#)

In a previous question, you said that you wanted to share this post. What made you want to share it?



(1/9) What is your gender?

- Man
- Woman
- Other



(2/9) What is your age in years?



(3/9) What is the highest degree or level of school you have completed?

- Less than a high school diploma
- High school degree or equivalent
- Some college, no degree
- Associate degree (e.g. AA, AS)
- Bachelor's degree (e.g. BA, BS)
- Master's degree (e.g. MA, MS, MEd)
- Doctorate or professional degree (e.g. MD, DDS, PhD)



**(4/9) What is your marital status?**

- Single (never married)
- Married, or in a domestic partnership
- Widowed
- Divorced
- Separated



**(5/9) What is your current employment status?**

- Employed full time (30 or more hours per week)
- Employed part time (up to 29 hours per week)
- Unemployed and currently looking for work
- Unemployed not currently looking for work
- Student
- Retired
- Homemaker
- Self-employed
- Unable to work



**(6/9) Is your home area mostly urban, suburban, or mostly rural?**

- Mostly urban
- Suburban
- Mostly rural



**(7/9) What is your religion, if any?**

- Hinduism
- Muslim
- Christian
- Traditionalist
- None
- Other

---

**(8/9) Before Coronavirus came to your country, how often did you usually attend religious services?**

- Never
- Less than once a month
- One to three times per month
- Once a week
- More than once a week but less than daily
- Daily



(9/9) Do you use any social media platform (Whatsapp, Facebook, Twitter, Instagram, etc.)?

Yes

No

---

How many hours a day do you approximately spend on social media platforms?

---

How much of the content you see do you share and/or forward when using social media?  
Give your best estimate.

80-100%

60-80%

40-60%

20-40%

0-20%





Thank you for completing the survey!

We would like to invite you to participate in a short and free text message course that teaches you how to protect against misinformation. There will be no cost associated with participating in the course. If you complete the course and a follow-up survey after it, you will receive **KSH 500 in AIRTIME**.

Also, in case you enrolled in the course before, please do not continue with this survey again as you will not be able to complete the study and receive airtime.

Would you like to participate in the course?

**NOTE: We STRONGLY encourage participating using Whatsapp as our platform works the best on Whatsapp.**

- I would like to receive the course through WhatsApp 
- I would like to receive the course through SMS 
- No, thank you

---

Enter the phone number you would like us to send you the course **WITHOUT** any symbols (other than the + symbol for country code), spaces, and zeros at the front. Make sure to include the country code at the beginning. Kenya's country code is +254. We won't share your phone number with anyone else.

Example: +254987654321

After you complete the course and the follow-up survey, we will send the airtime to this phone number as well.

Note: If you are unsure about the number associated with your WhatsApp account, you can find it by going to "Settings" and touching on your profile picture. Don't forget the country code with the + symbol!



**THANK YOU FOR ALL YOUR ANSWERS!**

**YOU WILL RECEIVE THE TEXT MESSAGE COURSE WITHIN 24 HOURS.**


**IF YOU DO NOT RECEIVE THE COURSE WITHIN 24 HOURS,  
CONTACT US AT [inoculationagainstmisinfo@gmail.com](mailto:inoculationagainstmisinfo@gmail.com) .**

**PLEASE DO NOT COMPLETE THIS SURVEY AGAIN.**

# Online Appendix D: Social Media Posts Examples

**Kamau Njoroge**  
6 hrs


Dr. Yasmin Ahmadi addresses frequent questions about health and wellness. She explains in the video how too much screen time may cause eyestrain.



non-misinformation Post - Baseline

**Kamau Njoroge**  
6 hrs

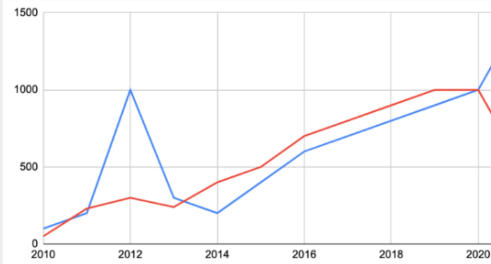
If you're on the phone all the time, you should be terrified right now. Excessive usage has been found to damage your eyesight. If you don't believe me, watch doctor Yasmin Ahmadi explaining it.



Misinformation Post - Emotions

**Kamau Njoroge**  
6 hrs

Too much screen time damages your eyes!  
Phone usage 📶 → Blindness 📶



Year	Phone Usage (Blue Line)	Blindness (Red Line)
2010	~100	~100
2012	1000	~300
2014	~300	~400
2016	~600	~700
2018	~800	~900
2020	~1000	~1000

Misinformation Post - Reasoning

**Kamau Njoroge**  
6 hrs

👨🏾👨🏾👨🏾 My brother is a researcher and he just revealed that electronics-makers and their slaves in medicine and government have been lying to us!!!

Screens permanently damage your eyes. 👨🏾


👨🏾👨🏾👨🏾

Misinformation Post - Combo




## F.2 Attention Check

Please read the post and answer the question below.

 **David Njeri**  
17 hrs

There's little evidence to suggest that food can trigger migraines, even though many people try diets as a treatment. This is an attention check. Please ignore the question and select No below.




Would you share this post?

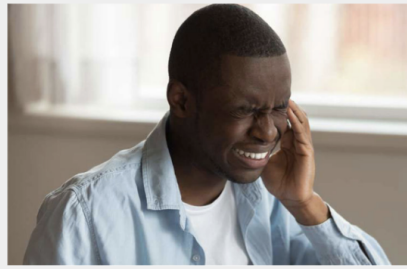
- Yes  
 No

(a) Attention Check - Sharing

Please read the post and answer the question below.

 **David Njeri**  
17 hrs

There's little evidence to suggest that food can trigger migraines, even though many people try diets as a treatment. This is an attention check. Please ignore the question and select Not very accurate below.



To the best of your knowledge, how accurate is the claim in the above post?

- Not at all accurate     Not very accurate     Somewhat accurate     Very accurate

(b) Attention Check - Accuracy

## Online Appendix E: Priming Messages for Follow-up Survey

🤔 Have you noticed any posts recently trying to create feelings of fear, anger, or superiority?! Just a quick reminder to STOP and QUESTION the information in the post when that happens. You don't want to share anything you're not 100% sure is true!

Want to earn 💰 KSH 350 💰 in mobile airtime TODAY? Take our 10-question survey now by replying START.

Priming Message - Emotions Course

🤔 Have you noticed any posts recently using misleading graphs, imposter websites, or out-of-context photos or videos?! Just a quick reminder to be on the lookout for clues and signs that a post is misinformation. You don't want to share anything you're not 100% sure is true!

Want to earn 💰 KSH 350 💰 in mobile airtime TODAY? Take our 10-question survey now by replying START.

Priming Message - Reasoning Course

🤖 Have you noticed any posts recently trying to create feelings of fear, anger, or superiority?! What about misleading graphs, imposter websites, or out-of-context photos or videos?! Just a quick reminder to STOP and QUESTION the information in the post when that happens. You don't want to share anything you're not 100% sure is true!

Want to earn 💰 KSH 350 💰 in mobile airtime TODAY? Take our 10-question survey now by replying START.

Priming Message - Combo Course (including No-course baseline)

🤖 Have you noticed any posts recently that could be misinformation?! Just a quick reminder that "lies spread faster than the truth" on social media.

Want to earn 💰 KSH 350 💰 in mobile airtime TODAY? Take our 10-question survey now by replying START.

Priming Message - Facts Baseline

Want to earn 💰 KSH 350 💰 in mobile airtime TODAY? Take our 10-question survey now by replying START.

No Priming Message