Leveraging Media Literacy Training to Promote Social Corrections

Catherine King[1 [0000−0002−1636−9887]] and Kathleen M. Carley[1 [0000−0002−6356−0238]]

Carnegie Mellon University, Pittsburgh PA 15213, USA
{cking2,kathleen.carley}@cs.cmu.edu

Abstract. This paper analyzes the effectiveness of utilizing traditional media literacy training in a new way: encouraging and training social media users to engage in social corrections or other countering behaviors online when they encounter misinformation in their news feeds. An experiment was run among government analysts, where participants were shown a series of false or misleading social media posts and asked to describe if and how they would respond to those posts. A survey was administered both before and after an interactive, in-person training session. After the training, respondents were more likely to claim they would intervene with more effort when seeing misinformation, for example, by employing social corrections either publicly by commenting on the post or privately by messaging the poster. However, that increase came primarily from individuals who were already claiming to engage in some low-effort interventions (like reporting misinformation) before the training rather than from individuals who were not engaging in any countering efforts at all. Finally, a qualitative analysis showed that the content of a post, the type of account that posted it, and the platform where it was posted all affect the willingness of the participants to engage in countering behavior. The promising results from this case study indicate that more work should be conducted in this domain.

Keywords: misinformation interventions · media literacy · social corrections

1 Introduction

As misinformation continues to affect societies around the world, there has been much recent research focused on developing and deploying effective interventions [4, 6, 8, 15]. One of the most studied intervention types involves media and digital literacy as a preventative measure [6]. Media literacy encompasses many different types of interventions, ranging from short tips [7], in-person training sessions [20], to fake news games [3, 16].

Most media literacy experiments focus on improving participants' skills so that they can better discern truth from falsehoods and enhance their critical understanding of the media they encounter [10]. While many studies also investigate the effectiveness of media literacy on behavioral outcomes, those studied behaviors typically focus on lowering the frequency of harmful, risky, or antisocial behaviors such as engaging with or sharing misinformation or partaking in risky sexual encounters. To our knowledge, no studies have yet focused on improving participants' willingness and ability to counter it [10]. We seek to fill this research gap by running an experiment focused on increasing motivated individuals' willingness, knowledge, and ability to counter misinformation. Our research questions:

1. Does targeted training increase the likelihood of countering misinformation?
2. What factors affect willingness to engage in interventions?

The results of this case study will help inform how to encourage and improve user-based countermeasures, such as social corrections, thereby contributing to broader efforts in combating misinformation.

2 Related Work

2.1 Media Literacy Interventions

Media literacy interventions consist of educational initiatives designed to enhance the public's civic discourse by improving critical thinking ability when reading media content [7, 10]. One type is the development and usage of fake news games. These games include the Bad News Game [3], Go Viral! [21], Troll Spotter [16], and Harmony Square [23]. They are designed to be an interactive and fun way to help players detect misinformation [19, 21].

A related concept to media literacy is the theory of inoculation, sometimes referred to as "pre-bunking". Inoculation includes interventions like preemptive warning messages or other anticipatory interventions meant to "inoculate" people, much like a vaccine would, from later believing that misinformation or harmful content [17]. Similarly, media literacy is also intended to improve participant's resilience when encountering misleading, harmful, or false messages.

The effectiveness of media literacy as a preventative measure against misinformation has been widely debated in the literature. There is a lack of consensus on whether it is effective, which types are effective, how effective it is, how long the effectiveness lasts, and in which contexts it is most effective [1, 3, 10, 21].

2.2 User-Based Interventions

In the context of countering misinformation, user-based interventions refer to actions that social media users can take when directly engaging with misinformation [2, 11]. For example, social media platforms typically allow users to report other users or posts [22]. Social media users can also employ social corrections. Social corrections attempt to debunk the misinformation poster by publicly commenting on their post, privately messaging the user, or other related means [2].

User-based measures are an essential type of misinformation countermeasure. While most platforms employ some automated moderation, they also rely on social media users to report anything those algorithms miss. Additionally, users can comment and engage in social corrections to help debunk the misinformation. Having a trusted messenger, such as a friend or family member, debunk misinformation has been found to be effective in several studies [2, 5, 18, 27, 29]. User-based interventions are a vital tool in the fight against misinformation, and there is currently little to no work being done utilizing media literacy interventions to improve those measures.

3 Methods

We ran an experiment testing the effectiveness of a countering training session on 23 government analysts. The participants had signed up for social cyber-security training and were located in Orlando, FL, at the time of the experiment. They were involved in a 2-week training exercise called "OMEN" (Operational Mastery of the Information Environment) [13], which ran from 02/05/24 - 02/16/24. The age of participants ranged from 21 to 58, with an average age of 35.6 and median age of 34 years old. Of the 23 analysts, 19 were men, and 4 were women.

3.1 Survey

Overview The survey was implemented in Qualtrics and had two sections. In the first section, participants were shown a set of 16 social media posts, randomly selected from a pool of posts, and asked a series of questions. These questions included asking what they thought the accuracy and trustworthiness of the posts were. In the second section, participants were shown four explicitly false posts and then asked what, if anything, they would do if they saw that post on their social media feeds. They were also asked if their answer would change depending on the poster of the message or the platform where they saw the message. For this paper, only the responses from the second section were analyzed. The responses from the first section are left for future work.

Post Selection We designed the survey posts to look like generic social media posts. The topics in the posts were chosen to be apolitical and timely, and included health (COVID-19, vaccines), science (climate change, flat earth theories), and recent entertainment topics. To control for the possible differences in the difficulty of assessing each post, we had a group of experts review the posts for both difficulty and quality. Between four and six reviewers were assigned to review each post. Based on reviewer comments, some posts were removed or modified. We used the average difficulty score to sort the remaining posts and then randomly split them into the pre and post-training surveys.

Responses to Seeing Misinformation When shown these false posts, participants could select from the list of possible responses shown in Table 1. This list of responses was developed in a previous paper [14] but was extended to include an “Other" option. Responses labeled as “Low Effort" apply to indirect or quick actions. Responses labeled as “High Effort" apply to actions that directly engage with the misinformation content and are more time-consuming. The participants who selected “Other" were prompted to write in their response, and the effort level of their response was manually reviewed.

Table 1. Actions one can take when engaging with misinformation on social media.

Response	**Effort Level**
Ignore the post	No Effort
Report the post	Low Effort
Report the user	Low Effort
Block the user	Low Effort
Unfollow or unfriend the user	Low Effort
Privately message the user	High Effort
Comment a correction on the post	High Effort
Create a separate post with the correct information	High Effort
Other	-

Textual Analysis of Open-Ended Questions The survey asked the participants to explain their reasoning about if and how the misinformation poster or the platform would affect their countering behavior. To analyze the text-based responses, we used code mapping to find common themes throughout the responses. Code mapping, sometimes known as affinity diagramming, is a common approach when analyzing open-ended survey data [9, 24].

We first read over all the responses per question, then sorted related comments into piles using Apple's Freeform program[1], which is a digital whiteboard. Whenever a participant wrote something like “same reason as my previous answer", we put those in the same pile as their previous related response. We iteratively generated codes by reviewing the responses a second time. Then, we tagged each group of responses with various factors and combined related factors into general themes. The major themes are analyzed in the Results section.

3.2 Countering Training

After administering the first survey, we gave a 30-minute interactive training session adapted from King's work on why and how to counter misinformation effectively [12]. After the training, participants were given the post-training survey. The training was broken down into three parts:

- *Why People Should Counter Misinformation* - The trainer and participants discussed common reasons why people do not counter misinformation and how those concerns can be addressed [25, 26].

[1] https://www.apple.com/newsroom/2022/12/apple-launches-freeform-a-powerful-new-app-designed-for-creative-collaboration/

– *Common Logical Fallacies* - The trainer reviewed logical fallacies and how to spot them. Examples came from several university research guides[2][3][4].
– *Effective Interventions* - Finally, participants were trained on types of interventions and debunking efforts that are effective [18, 28].

3.3 Ethics Information

The Carnegie Mellon University Institutional Review Board (IRB) approved this study, numbered "STUDY2023_00000429". They determined the study to be exempt from a full review because it involved a “benign behavioral intervention". All participants were randomly assigned a user ID to link their pre and post-training results, but their names were not collected. The survey collected informed consent from all participants. Participants were not paid by the study but were instead paid their typical government salary.

4 Results and Analysis

4.1 Countering Responses

We found that the maximum amount of effort participants selected to counter any of the misinformation posts increased from the pre-training survey to the post-training survey (see Table 2). More people said they would engage in high-effort actions, and this increase came from people already engaging in low-effort actions.

Table 2. The percentage and number of participants whose maximum effort level was in each of the three effort level categories described in Table 1.

	Pre-Training	**Post-Training**
No Effort	30.4% (7)	34.8% (8)
Low Effort	65.2% (15)	39.1% (9)
High Effort	4.3% (1)	26.1% (6)

4.2 Factors Affecting Countering Actions

For each post, participants were asked if their answer would change depending on the person or organization posting it or on which platform they saw the misinformation. Figure 1 shows the total number of times participants replied with each possible answer over both surveys. “Probably Not" and “Definitely Not" were selected the most frequently for both poster and platform. However, over

[2] https://libguides.princeton.edu/c.php?g=982190&p=7102155
[3] https://owl.purdue.edu/owl/general_writing/academic_writing/
[4] https://writingcenter.unc.edu/tips-and-tools/fallacies/

all the posts shown, 18 participants (78.3%) and seven participants (30.4%) answered “Probably Yes" or “Definitely Yes" for at least one post when asked if the poster or the platform respectively would change their answer. Participants were more likely to say that the misinformation poster would affect their countering response than the platform.

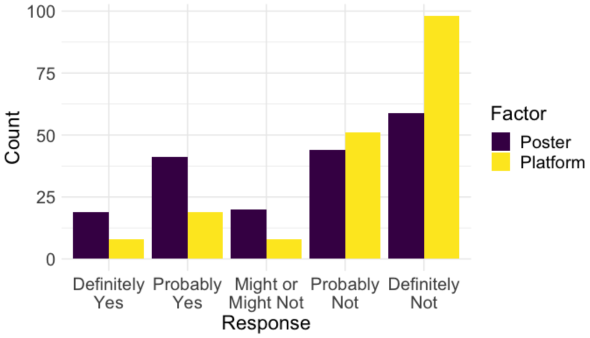

Fig. 1. The number of times participants said their answer would change depending on either the platform or the poster over both the pre and post-training surveys.

When elaborating upon their responses on how these factors would affect their countering efforts, we found four main recurring themes. Figure 2 shows the most commonly mentioned sentiments among those themes.

Theme 1. Platform and Account Preference: Many participants prefer engaging with close contacts and engaging on specific platforms more than others. Most participants (56.2%) mentioned that if they knew the poster of the misinformation, they would be more likely to engage in direct debunking efforts, whether on the social media post, through a private message, or offline. On the other hand, most participants either did not mention anything to do with platform preferences (10, or 43.5%) or said that they always treat all platforms equally (11, or 47.8%), while only two respondents (8.7%) mentioned having a platform preference. For example, one participant stated: “I am not likely to engage with users on sites where my identity is directly tied to the account. Accounts like Reddit, where I am more anonymous, makes discussion easier to partake in and exit from." Overall, 17 participants (73.9%) mentioned preferences in some way across both surveys and factors.

Theme 2. Content: The post’s content was one of the most commonly mentioned reasons for either engaging or not engaging in countering efforts. Twenty respondents (87%) mentioned content at least once when elaborating on either platform or poster reasoning. Specifically, if participants perceived the content as extreme or incredulous, many said they thought any effort would be wasted. For example, one respondent mentioned, “I just don’t think i [sic] could change the mind of someone who believed the earth was flat." Conversely, if

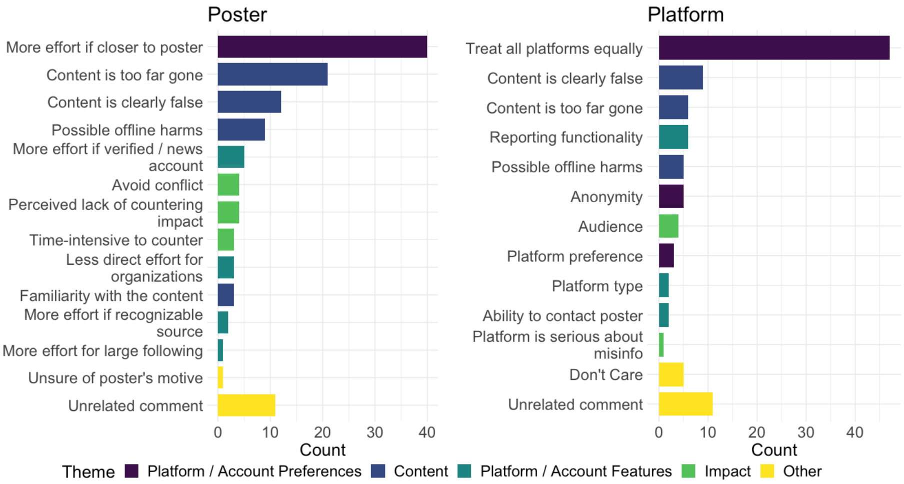

Fig. 2. This figure displays the total number of times over both surveys that participants mentioned a specific sub-theme when asked if and how their answer would change depending on the misinformation poster and the platform.

something was easy to debunk, like a straightforward or factual error, a topic they knew a lot about, or if the post had the potential for severe offline consequences, participants mentioned they would be more likely to engage.

Theme 3. Platform and Account Features: Four participants (17.4%) mentioned the importance of account features and said they would be more likely to intervene if the poster was either a verified account, a news agency, a recognizable source, or had a large following. While many participants mentioned they treat all their social media platforms equally, three participants (13%) mentioned that they thought they would be more likely to respond on some platforms than others based on features of that platform. For example, one participant stated: “If I were in a channel specific to that topic, I may look to validate/invalidate the content. But on a platform like facebook [sic] I would be less likely to give it a second look." The ease with which one could report a post or contact the misinformation poster was also mentioned.

Theme 4. Impact: Seven participants (30.4%) mentioned the potential impact of countering, and most used it as a reason why they were unlikely to take action. Participants mentioned factors like that they felt that it would be too time-intensive to debunk, would have little to no impact, did not think they knew enough to counter the post, or wanted to avoid conflict. One participant mentioned impact positively and said that taking action on some platforms that take moderation seriously may be more impactful than others with less moderation. However, another participant felt the opposite: “I would be more likely to report it on an application or site with worse media literacy."

5 Discussion and Conclusions

We conducted an experiment on the effectiveness of media literacy training to increase the willingness and likelihood of social media users utilizing countering actions. We found that after the training, more respondents on the survey claimed they would intervene more directly (see Table 2). This increase came primarily from individuals already engaging in low-effort countering efforts, such as reporting or blocking.

We qualitatively analyzed participants' explanations when asked if and how their likelihood of countering would change depending on the account posting the misinformation and the platform on which it was posted. Overall, we found that even though the **Content** theme was mentioned by the most number of unique participants (20), **Platform and Account Preferences** were the most frequently mentioned factors across all posts. As seen in Figure 1, closeness to poster and platform neutrality heavily dominate all other factors. These results likely indicate that individuals with these beliefs feel them strongly and across posts. Other frequently mentioned themes were **Platform and Account Features** and possible **Impact** (or lack of impact) when countering.

There are multiple limitations to this work. First, the sample size is relatively small. While this was helpful in order to get detailed qualitative feedback and analysis, it does indicate that more future work is needed in this area to confirm these results. Second, the participants were government analysts and were more educated than the average American (all participants had at least some college, with most having bachelor's degrees or even higher). However, participants were motivated because they continued to get their full-time salary if they took this training, and they received training credits. This led to very detailed and thoughtful qualitative responses.

Understanding why people do or do not counter and showing that training can increase people's willingness to intervene is crucial in determining how to improve social corrections and other user-based countermeasures in the future. For example, closeness was a significant factor. People may feel more comfortable with those closer to them and feel like they can make more of a difference or should at least try among loved ones. Verified accounts or those with large followings were considered more important to counter. The ease with which one could report a post or user was also mentioned. Knowing these factors, social media companies can better design their platforms. For example, platforms can encourage more reporting by improving reporting functions and can encourage more social corrections by showing more posts from closer contacts.

Acknowledgments. This work was supported in part by the Office of Naval Research (ONR) Award N000142112229 for "Scalable Tools for Social Media Assessment", the Knight Foundation, the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the Center for Informed Democracy and Social-cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government. We would also like to thank Christine Sowa Lepird for her involvement in Project OMEN.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Badrinathan, S.: Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India. American Political Science Review **115**(4), 1325–1341 (Nov 2021). https://doi.org/10.1017/S0003055421000459
2. Badrinathan, S., Chauchard, S.: "I Don't Think That's True, Bro!" Social Corrections of Misinformation in India. The International Journal of Press/Politics p. 19401612231158770 (Feb 2023). https://doi.org/10.1177/19401612231158770
3. Basol, M., Roozenbeek, J., van der Linden, S.: Good News about Bad News: Gamified Inoculation Boosts Confidence and Cognitive Immunity Against Fake News. Journal of Cognition **3**(1), 1–9 (2020). https://doi.org/10.5334/joc.91
4. Blair, R.A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., Stainfield, C.J.: Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. Current Opinion in Psychology **55**, 101732 (Feb 2024). https://doi.org/10.1016/j.copsyc.2023.101732
5. Bode, L., Vraga, E.K.: See Something, Say Something: Correction of Global Health Misinformation on Social Media. Health Communication **33**(9), 1131–1140 (Sep 2018). https://doi.org/10.1080/10410236.2017.1331312
6. Courchesne, L., Ilhardt, J., Shapiro, J.N.: Review of social science research on the impact of countermeasures against influence operations. Harvard Kennedy School Misinformation Review (Sep 2021). https://doi.org/10.37016/mr-2020-79
7. Guess, A.M., Lerner, M., Lyons, B., Montgomery, J.M., Nyhan, B., Reifler, J., Sircar, N.: 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 (Jul 2020). https://doi.org/10.1073/pnas.1920498117
8. Helmus, T.C., Kepe, M.: A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda. Tech. rep., RAND Corporation (Jun 2021), https://www.rand.org/pubs/research_reports/RRA894-1.html
9. Jackson, K.M., Trochim, W.M.K.: Concept Mapping as an Alternative Approach for the Analysis of Open-Ended Survey Responses. Organizational Research Methods **5**(4), 307–336 (Oct 2002). https://doi.org/10.1177/109442802237114
10. Jeong, S.H., Cho, H., Hwang, Y.: Media Literacy Interventions: A Meta-Analytic Review. The Journal of Communication **62**(3), 454–472 (Jun 2012). https://doi.org/10.1111/j.1460-2466.2012.01643.x
11. Jhaver, S., Zhang, A.X.: Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. New Media & Society p. 14614448231217993 (Dec 2023). https://doi.org/10.1177/14614448231217993
12. King, C.: Thesis Proposal: Effective and Practical Strategies for Combatting Misinformation. Ph.D. thesis, Carnegie Mellon University
13. King, C., Lepird, C.S., Carley, K.M.: Project OMEN: Designing a Training Game to Fight Misinformation on Social Media. Tech. rep. (2021), http://reports-archive.adm.cs.cmu.edu/anon/isr2021/abstracts/21-110.html
14. King, C., Phillips, S.C., Carley, K.M.: Registered Report: A path forward on online misinformation mitigation based on current user behavior. Scientific Reports (2024), stage 1 Manuscript, Forthcoming

15. Kozyreva, A.e.a.: Toolbox of individual-level interventions against online misinformation. Nature Human Behaviour pp. 1–9 (May 2024). https://doi.org/10.1038/s41562-024-01881-0
16. Lees, J., Banas, J.A., Linvill, D., Meirick, P.C., Warren, P.: The Spot the Troll Quiz game increases accuracy in discerning between real and inauthentic social media accounts. PNAS Nexus **2**(4), pgad094 (Apr 2023). https://doi.org/10.1093/pnasnexus/pgad094
17. Lewandowsky, S., van der Linden, S.: Countering Misinformation and Fake News Through Inoculation and Prebunking. European Review of Social Psychology **0**(0), 1–38 (Feb 2021). https://doi.org/10.1080/10463283.2021.1876983
18. Lewandowsky, S.e.a.: Debunking Handbook 2020. Tech. rep., Databrary (2020), http://databrary.org/volume/1182
19. Maertens, R., Roozenbeek, J., Basol, M., van der Linden, S.: Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. Journal of Experimental Psychology: Applied **27**(1), 1–16 (2021). https://doi.org/10.1037/xap0000315
20. McGrew, S.: Learning to evaluate: An intervention in civic online reasoning. Computers & Education **145**, 103711 (Feb 2020). https://doi.org/10.1016/j.compedu.2019.103711
21. Modirrousta-Galian, A., Higham, P.A.: Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. Journal of Experimental Psychology: General **152**(9), 2411–2437 (2023). https://doi.org/10.1037/xge0001395
22. Niklewicz, K.: Weeding Out Fake News: An Approach to Social Media Regulation. European View **16**(2), 335–335 (Dec 2017). https://doi.org/10.1007/s12290-017-0468-0
23. Roozenbeek, J., Linden, S.v.d.: Breaking Harmony Square: A game that "inoculates" against political misinformation. Harvard Kennedy School Misinformation Review (Nov 2020). https://doi.org/10.37016/mr-2020-47
24. Saldaña, J.: The Coding Manual for Qualitative Researchers. Sage, 2nd edn. (2013)
25. Southwell, B.G., Thorson, E.A., Sheble, L.: Misinformation and mass audiences
26. Tandoc, E.C., Lim, D., Ling, R.: Diffusion of disinformation: How social media users respond to fake news and why. Journalism **21**(3), 381–398 (Mar 2020). https://doi.org/10.1177/1464884919868325
27. Tay, L.Q., Hurlstone, M.J., Kurz, T., Ecker, U.K.H.: A comparison of prebunking and debunking interventions for implied versus explicit misinformation. British Journal of Psychology **113**(3), 591–607 (2022). https://doi.org/10.1111/bjop.12551
28. Trethewey, S.P.: Strategies to combat medical misinformation on social media. Postgraduate Medical Journal **96**(1131), 4–6 (Jan 2020). https://doi.org/10.1136/postgradmedj-2019-137201
29. Walter, N., Brooks, J.J., Saucier, C.J., Suresh, S.: Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis. Health Communication **36**(13), 1776–1784 (Nov 2021). https://doi.org/10.1080/10410236.2020.1794553