# A field experiment to mitigate the harm of inaccurate information online

[1]Giovanni Luca Ciampaglia, [1]Do Won Kim, [2]Brendan Nyhan, [3]Ro'ee Levy, [4]Betsi Grabe, [5]Filippo Menczer

*[1]University of Maryland, [2]Dartmouth College, [3]Tel Aviv University, [4]Boston University, [5]Indiana University*

## Introduction

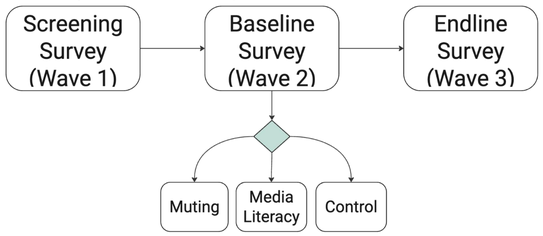
**Objective:** Develop cost-effective, sustainable, scalable interventions to combat misinformation.

## Research Questions

1. What are effective interventions to achieve such a goal?
2. What are the effects of such interventions on behavioral and attitudinal outcomes?

## Method

We piloted a two-arm, three-wave RCT on Twitter/X (treatment period: 1 month).




Arm 1. Muting low-quality sources

- Target information environments to increase costs of getting low-quality information

Arm 2. Media literacy tips

- Target individual discernment to reduce decision costs of assessing information quality

## Outcome variables

- Engagements w/ low-quality sources (likes, retweets, etc.)
- Sharing intention and accuracy judgement of True / False headlines (drawn from Fazio et al., 2024)
- Trust in different info sources, etc.

## References

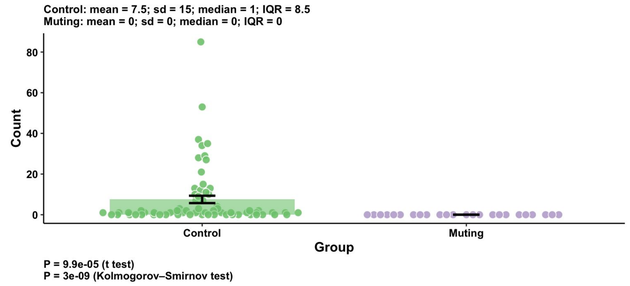
Fazio, L., Rand, D. G., Lewandowsky, S., Susmann, M., Berinsky, A. J., Guess, A. M., … Swire-Thompson, B. (2024, June 23). Combating misinformation: A megastudy of nine interventions designed to reduce the sharing of and belief in false and misleading headlines. https://doi.org/10.31234/osf.io/uyjha

# Encouraging **Muting** and Giving **Media Literacy** Tips Decreases **Engagement** with Low-quality **News** Sources on **Twitter** / X

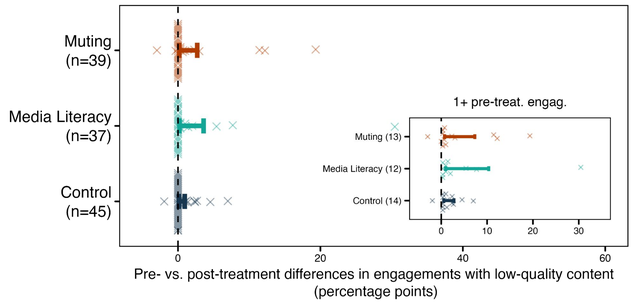
For more information, please contact dowonkim@umd.edu or scan the QR code.

## Results (Pilot Study)

Manipulation Check: Muting will create sustained changes to information exposure (i.e. reduce post-treatment exposure to muted low-quality sources)




Pilot study randomized whether Twitter / X users (*n*=91) were offered incentives to either mute low-quality sources or receive media literacy tips. The inset shows only participants with 1+ pre-treatment likes/(re)tweets/quotes.




- X-axis: changes in the proportion of engagements with low-quality sources relative to total engagements.
  - 0 indicates no change in engagement rate before and after treatment.
  - Positive values indicate a decrease in engagement rate post-treatment. E.g., 20 = proportion of engagements with low-quality accounts decreased by 20 percentage points after treatment.
- Error bars are 95% bootstrapped CIs around the mean difference in engagement rates.

## Going forward

Improve our targeted recruitment (i.e., people with low-quality information diets; consumers of untrustworthy news sources and misinformation superspreaders).

## Implications

- If muting proves more effective: Policy focus should be on platform-level interventions (e.g., banning or down-ranking low-quality accounts).
- If media literacy tips are more effective: Emphasis should be on user education programs for digital news consumption.

## Acknowledgements

This project is part of the Mercury Project (Social Science Research Council) and funded by the Sloan Foundation.