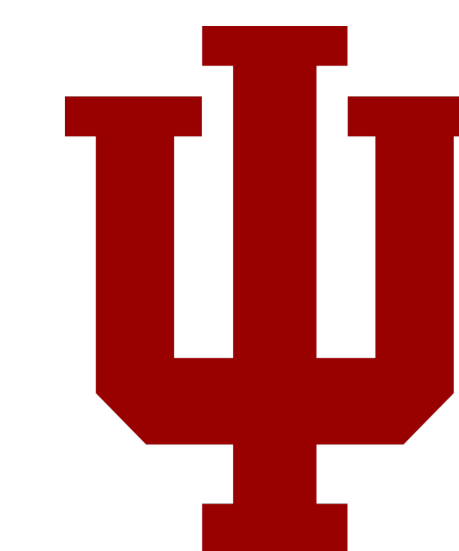# Boosting Individuals to Reduce Decisions to Engage with Distorted content Online

Eeshan Hasan   Gunnar Epping   Lorenzo Lorenzo-Luaces   Johan Bollen   Jennifer S. Trueblood

Indiana University Bloomington

Introduction

We examine whether a behavioral intervention based on boosting can provide individuals with the necessary cognitive tools to navigate social media, especially given its potential negative impacts on individual mental health.

- What is the impact of training individuals to identify exaggerated, emotional and unrealistic ways of thinking (called cognitive distortions) in online content?
- Does it help vulnerable populations such as depressed individuals?

Experimental Design

The stimuli were generated using an LLM, validated by a licensed clinical psychologist and ensured to be of similar sentiment. To maintain ecological validity, participants interacted with a realistic Twitter interface. The order of the training-identification block and interaction blocks was counterbalanced to study the impact of training. The depression severity was measured using the PHQ-9 scale.
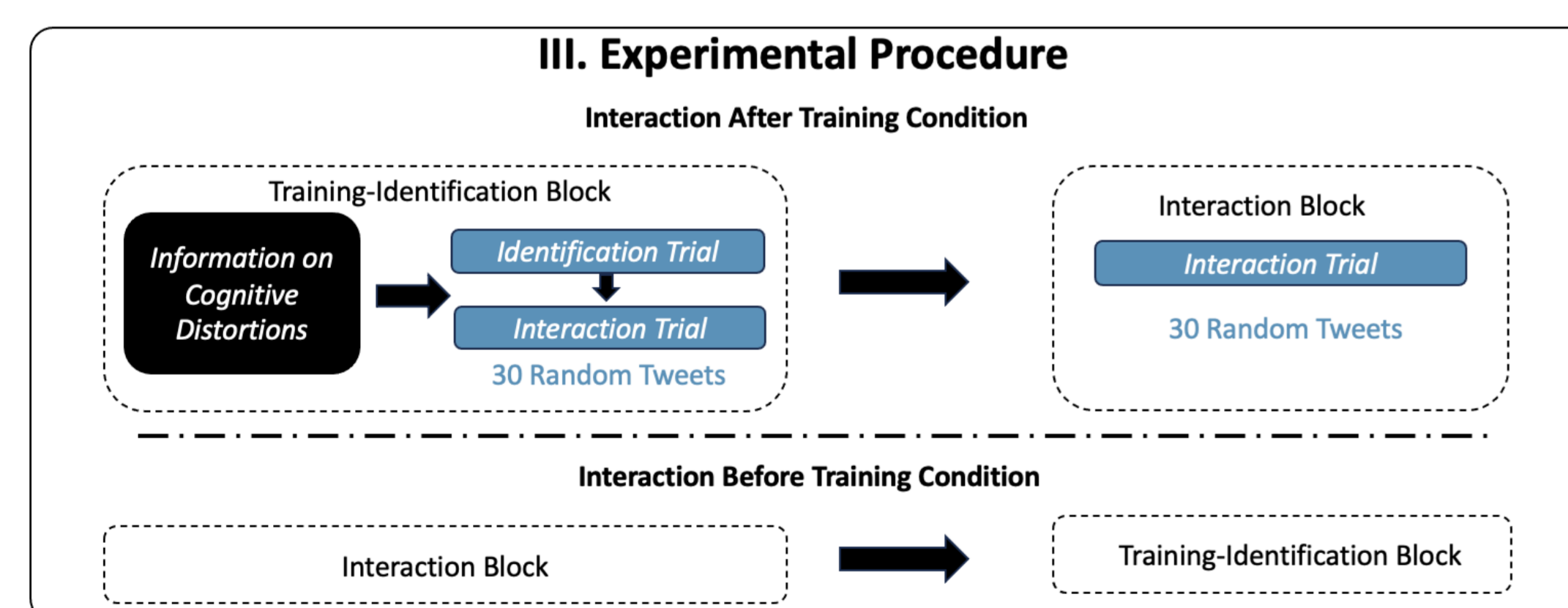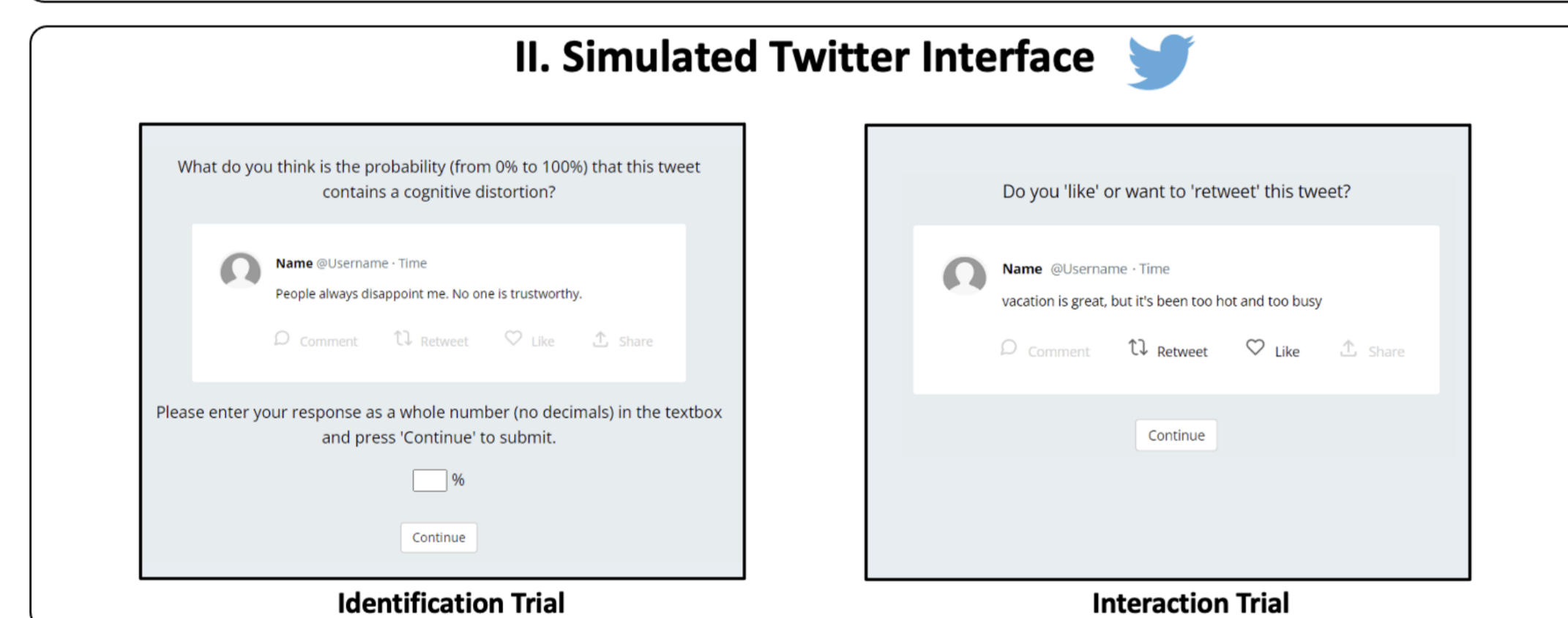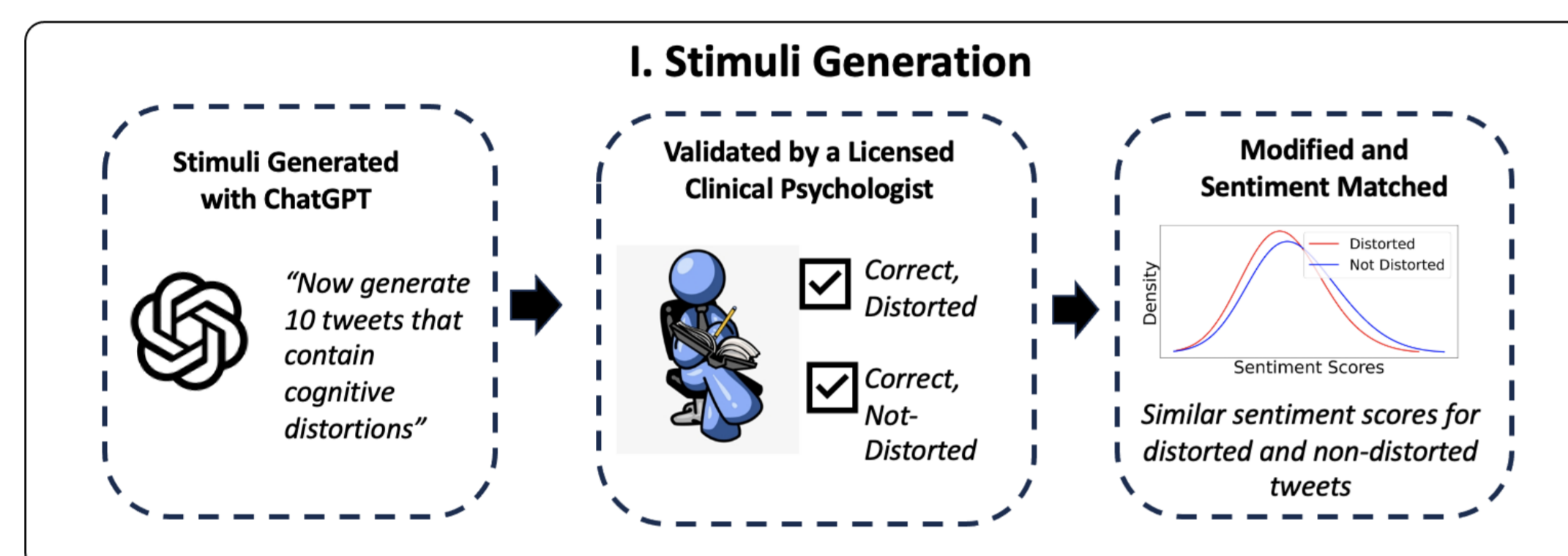
The methods used generate the stimuli, design interface and experimental procedure.

Participants

A total of 1000 individuals were recruited on MTurk using the platform Cloud Research from the United States. After exclusions (36 incomplete , 116 attention check, 10 invalid entries) 838 participants were retained.

Measures of Interest

- **Accuracy:** Mean correct (Binarized at 50%)
- **Slope:** Mean Judgment on Distorted Items - Mean Judgment on Non-Distorted Items.
- **Like (Retweet) Rate:** The rate of liking (retweeting) in the interaction block.
- **Liked (Retweet) Ratio Distorted/Total Liked (Retweeted):** Fraction of distorted liked by total liked (retweeted).
- **Depression Severity:** A clinically validated depression severity scale - PHQ-9.
- **Twitter Use Score:** Composite score based on time spent, active (vs passive) use and willingness to pay.
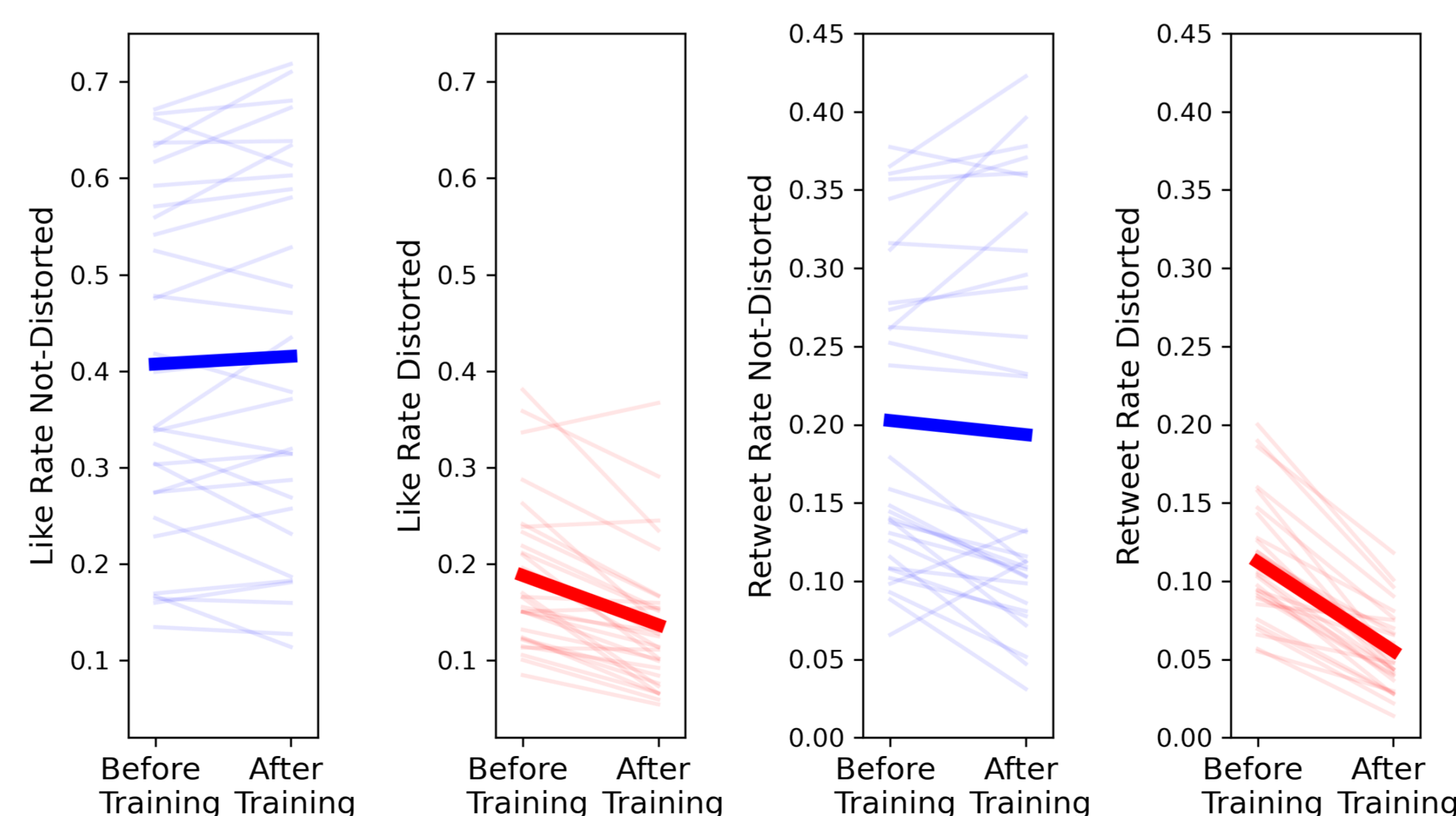
Correlation Results
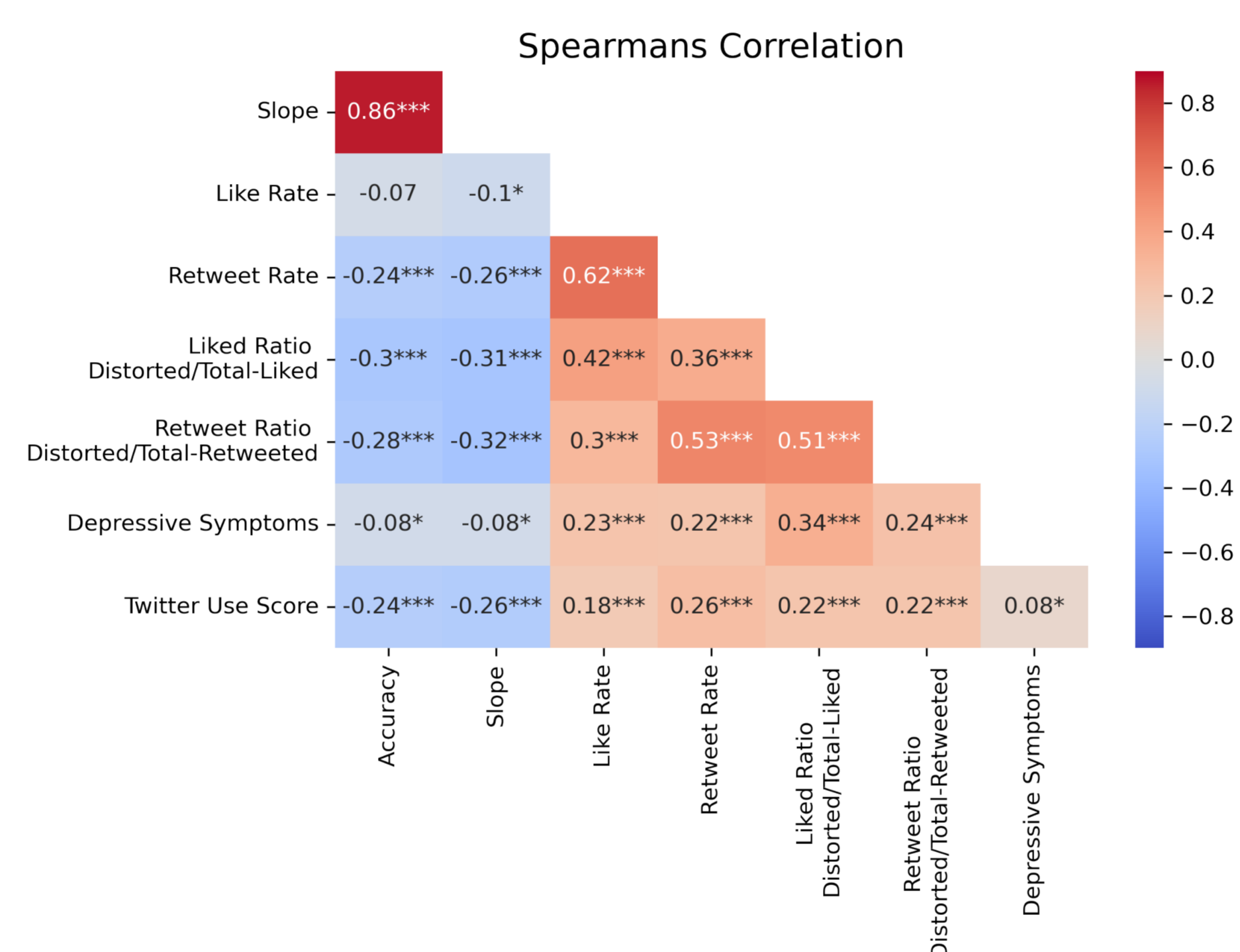
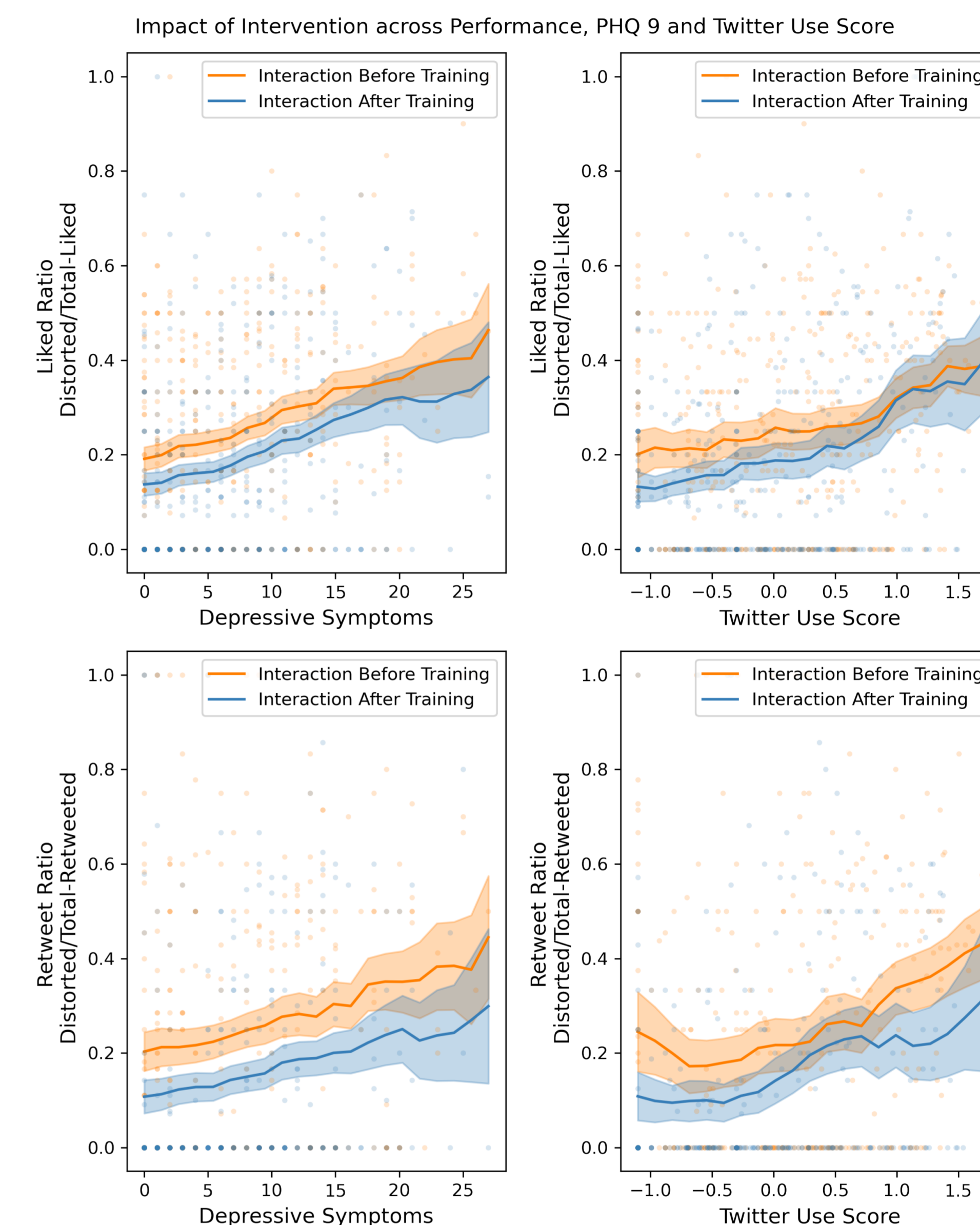Correlation of the individual differences among measures of interest.

Impact of Intervention

The training intervention reduced the propensity to like and retweet distorted items but had no impact on non-distorted items.

Impact of Intervention Across Participants

The intervention reduces the propensity to like and retweet content across the depression scale.

Discussion

- Individuals with greater depression severity interacted with distorted content more than healthier individuals.
- A psycho-educational micro-intervention that trained individuals to detect cognitive distortions was very effective in teaching individuals to distinguish distorted language from non-distorted language.
- The training sharply reduced the like rate for distorted content by around 30% and the retweet rate for distorted content by around 50%.
- This intervention reduced the sharing propensity across the depression severity scale.

References & Acknowledgments

Bathina, K. C., Ten Thij, M., Lorenzo-Luaces, L., Rutter, L. A., & Bollen, J. (2021). Individuals with depression express more distorted thinking on social media. Nature human behaviour, 5(4), 458-466.

Bollen, J., Ten Thij, M., Breithaupt, F., Barron, A. T., Rutter, L. A., Lorenzo-Luaces, L., & Scheffer, M.(2021). Historical language records reveal a surge of cognitive distortions in recent decades. Proceedings of the National Academy of Sciences, 118(30), e2102061118.

Herzog, S. M., & Hertwig, R. (2024). Boosting: Empowering citizens with behavioral science. Annual Review of Psychology, 76.