Title: Supplementary methods & analyses appendix for "Breaking *Harmony Square*: A game that "inoculates" against political misinformation"
Authors: Jon Roozenbeek (1), Sander van der Linden (1)
Date: November 6th, 2020
Note: The material contained herein is supplementary to the article named in the title and published in the Harvard Kennedy School (HKS) Misinformation Review.

# Appendix: Supplementary methods & analyses

*Supplementary information on sample selection and study participants*

For this study, we recruited participants via the online platform Prolific Academic (Palan & Schitter, 2018; Peer et al., 2017). Based on previous research (Basol et al., 2020), we first conducted an a priori power analysis using G* power, with $\alpha = 0.05$, $f = 0.26$ ($d = 0.52$), power of 0.95, and 2 experimental conditions. The minimal sample size required for detecting the main effect was 258. In total, 681 people were recruited in 2 separate data collections; a US-only sample ($n = 312$) and an international sample ($n = 369$). We pooled the results here (effect-sizes are slightly larger for the US-only sample). In total, 296 participants played *Harmony Square* (the treatment group), and 385 people played *Tetris* (the control group). This discrepancy is explained by the fact that we only included participants in the treatment group that played through the game in its entirety; following quality-control practices from previous research (Basol et al., 2020; Maertens et al., 2020). Specifically, participants in the treatment group were required to fill in a code before proceeding to the next stage of the study, which only appeared after finishing the game. Some participants ($n = 78$) entered the wrong code or no code at all, and were thus excluded from the dataset[1]. This is important because otherwise we cannot ensure that participants played through the whole game. No other exclusions were applied.

In total, 46.3% of our participants were from the United States, 11.0% from Portugal, 10.9% from Poland, 6.8% from the United Kingdom, 5.6% from Italy, 4.1% from Mexico, and another 15.4% from elsewhere. 43.2% of participants identified as female, 55.7% as male, and 1.2% as other (e.g. non-binary or agender). Participants were mostly younger, with 41.4% being between 18 and 24 years of age. The average education level was high, with 62.4% of participants indicating that they have at least a Bachelor's degree. The sample also skewed somewhat left in terms of political ideology, with the average score on the 1-7 political ideology scale (1 being "very left-wing" and 7 being "very right-wing") $M = 3.13$, $SD = 1.44$. On average, participants were paid £2.42 (or US $3.12). The average completion time was around 20 minutes. Supplementary Table S1 gives a detailed overview of the sample that was recruited for this study; it also shows that the sample of individuals that did not enter the correct completion code after playing *Harmony Square* and were thus excluded ($n = 78$) does not differ meaningfully from the rest of the sample, aside from their political ideology (which skews slightly more to the right for excluded participants).

*Supplementary analyses & robustness checks*

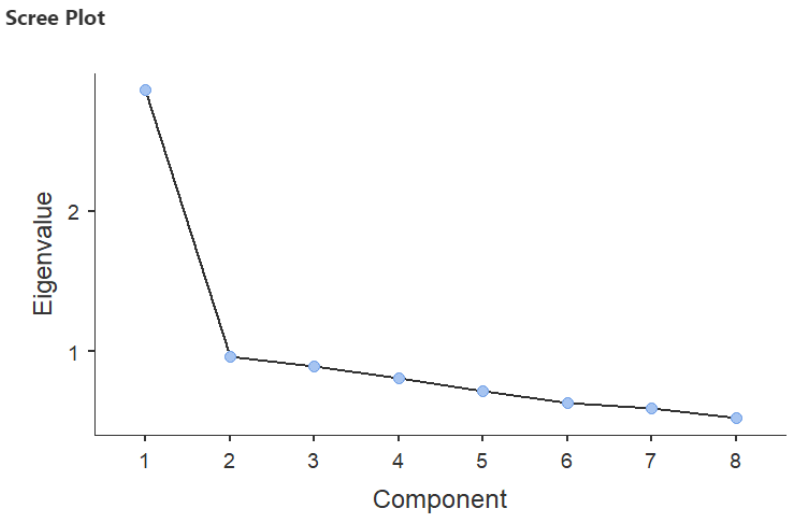
We conducted two separate robustness checks to validate the main analyses. First, we ran a linear regression analysis for each of the 3 outcome variables above, with the post-test as the dependent (outcome) variable, the condition (control or treatment) as a dummy variable, and the pre-test as the independent variable, for the reliability judgments, confidence judgments, as well as participants' willingness to share manipulative content. This analysis gives the same result as the ANOVA analysis that we ran for the difference scores above. The linear regression models for each outcome variable can be found in Supplementary Table S4. Second, following Pennycook et al. (2020), we also conducted a multi-level analysis with robust standard errors at the rating level, clustered on study participants and all 16 items (pre- and post-intervention). We find a significant interaction between pre-post differences and the

[1] The dataset for the excluded participants ($n = 78$) is available on the OSF: https://osf.io/r89h3/.

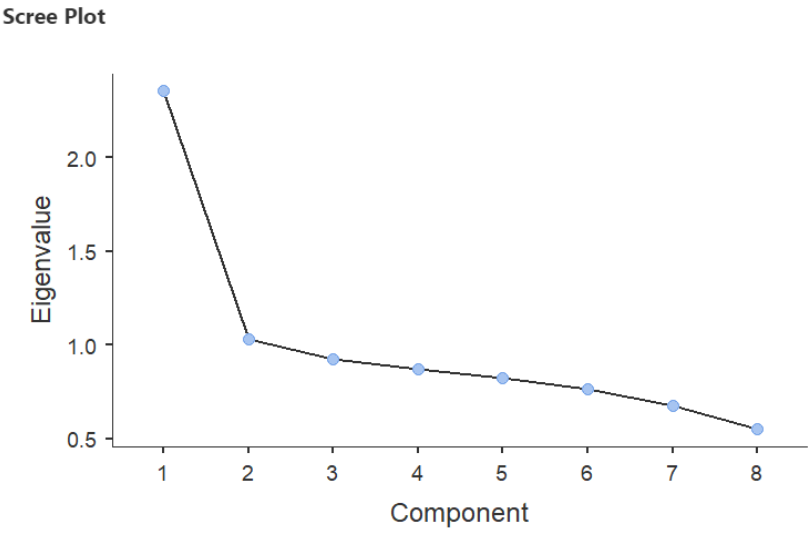
treatment (inoculation) condition for the reliability, confidence and sharing measures, further validating the results reported above. These results are reported in Supplementary Table S5.

*Items (social media posts) selection procedure & Principal Component Analysis*

To maintain balance, we selected 4 posts per manipulation technique (2 fictional and 2 "real"), for a total of 2 sets of 8 items (16 items in total). We conducted an exploratory principal component analysis (PCA) on both the "real" and the "fictional" item sets. Both sets loaded on a single dimension, with an eigenvalue of 2.35 for the "real' item set (accounting for 29.4% of the variance), and 2.86 for the "fictional" item set (accounting for 35.7% of the variance). Thus, for ease of interpretation and to limit multiple testing, both item sets were collapsed and treated as two measures, which we report throughout the paper. See Supplementary Figures S2 and S3 for the scree plots. To check for technique-level results, we also report the results for each individual manipulation technique taught in the game (both for the "real" and the fictional misinformation items) in Supplementary Table S3. In addition, descriptive statistics for each of the 16 items can be found in Supplementary Table S2.




***Supplementary Figure S2. Scree plot for reliability judgments following PCA for the "fictional" misinformation items.***




***Supplementary Figure S3. Scree plot for reliability judgments following PCA for the "real" misinformation items.***