Journal of Trial and Error

Volume 3
Issue 1
December 18, 2023

ISSN 2667-1204
https://doi.org/10.36850/i3.1

Contents

No Meaningful Difference in Attentional Bias Between Daily and Non-Daily Smokers
by *James Bartlett, Rebecca Jenks, and Nigel Wilson*

Gamified Inoculation Against Misinformation in India: A Randomized Control Trial
by *Trisha Harjani, Melisa-Sinem Basol, Jon Roozenbeek, and Sander van der Linden*

Driven to Snack: Simulated Driving Increases Subsequent Consumption
by *Floor van Meer, Stephen Murphy, Wilhelm Hofmann, Henk van Steenbergen, and Lotte F. van Dillen*

Challenges of Using Signaling Data From Telecom Network in Non-Urban Areas
by *Håvard Toft, Alexey Sirotkin, Markus Landrø, Rune Verpe Engeset, and Jordy Hendrikx*

Empathic Accuracy, Mindfulness, and Facial Emotion Recognition: An Experimental Study
by *Marije aan het Rot, Merle-Marie Pittelkow, D. Elisabeth Eckardt, Nils Simonsen, and Brian D. Ostafin*

An Introduction to Complementary Explanation
by Joeri van Hugten

No Meaningful Difference in Attentional Bias Between Daily and Non-Daily Smokers

James Bartlett[1], Rebecca Jenks[2], Nigel Wilson[3]

Both daily and non-daily smokers find it difficult to quit smoking long-term. One factor associated with addictive behavior is attentional bias, but previous research in daily and non-daily smokers found inconsistent results and did not report the reliability of their cognitive tasks. Using an online sample, we compared daily (n = 106) and non-daily (n = 60) smokers in their attentional bias towards smoking pictures. Participants completed a visual probe task with two picture presentation times: 200ms and 500ms. In confirmatory analyses, there were no significant effects of interest, and in exploratory analyses, equivalence testing showed the effects were statistically equivalent to zero. The reliability of the visual probe task was poor, meaning it should not be used for repeated testing or investigating individual differences. The results can be interpreted in line with contemporary theories of attentional bias where there are unlikely to be stable trait-like differences between smoking groups. Future research in attentional bias should focus on state-level differences using more reliable measures than the visual probe task.

Keywords *attentional bias, daily smokers, non-daily smokers, visual probe task, equivalence testing*

[1]School of Psychology and Neuroscience, University of Glasgow

[2]School of Psychological, Social and Behavioural Sciences, Coventry University

[3]School of Psychology and Sociology, Arden University

Received
December 13, 2021
Accepted
November 15, 2022
Published
December 13, 2022
Issued
December 18, 2023

Correspondence
School of Psychology and Neuroscience, University of Glasgow
james.bartlett@glasgow.ac.uk

Historically, smokers have been treated as a single homogeneous group (Shiffman, **2009**), but there are fundamental differences in the smoking habits and motives of daily and non-daily smokers (Shiffman, Dunbar, et al., **2012**; Shiffman, Tindle, et al., **2012**). Non-daily smokers make up 13 to 36% of smokers across Europe and the United States (Bogdanovica et al., **2011**; Kotz et al., **2012**; Tindle & Shiffman, **2011**) and non-daily smoking has typically been the most prevalent pattern in ethnic minority groups (Fagan & Rigotti, **2009**; Tong et al., **2006**). Whereas daily smokers cite negative reinforcers such as avoiding nicotine withdrawal as the key motivators, non-daily smokers cite positive reinforcers such as smoking around friends and alcohol (Shiffman, Dunbar, et al., **2012**; Shiffman et al., **2014**). Despite these differences, 77 to 92% of daily smokers and 74 to 83% of non-daily smokers relapse within 90 days of an attempt to quit (Bogdanovica et al., **2011**; Kotz et al., **2012**; Tindle & Shiffman, **2011**), showing both groups find it difficult to quit smoking long-term. This means it is important to investigate potential factors associated with smoking behavior.

One factor is *attentional bias*, which is the tendency to fixate on cues associated with smoking. Attentional bias is the product of a classical conditioning process where smokers develop conditioned responses to substance-related cues through repeated exposure (Field & Cox, **2008**). Theoretical models of attentional bias suggest it has a reciprocal relationship with craving and are supported by a meta-analysis showing there is a small positive relationship (Field et al., **2009**). In situations where cigarettes are available, cues associated with smoking grab attention and induce craving, which further drives attentional bias. Updated theories of attentional bias emphasize the role of momentary evaluations of smoking cues, meaning the levels of attentional bias and craving fluctuate over time, and describe attempts to extinguish the conditioned response through attentional bias modification (Field et al., **2016**). Consistent with theoretical models that suggest attentional bias is the result of

Take-home Message

In previous research using the visual probe task, some studies found that attentional bias towards smoking cues was greater in daily smokers, while others found attentional bias was greater in non-daily smokers. In our study, we found no meaningful difference using the traditional approach of analyzing differences in response times. Our visual probe task also showed poor reliability, meaning response time outcomes from the task should not be used when studying individual differences or measuring changes in attentional bias across repeated measurements.

repeated exposure to smoking cues, smokers as a single group consistently show greater attentional bias towards smoking cues than non-smokers (Baschnagel, **2013**; Ehrman et al., **2002**; Kang et al., **2012**; Mogg et al., **2003**). However, there are contrasting expectations and findings on how lighter and heavier smokers[1] differ in attentional bias.

On the one hand, lighter smokers should show greater attentional bias than heavier smokers since they rarely show signs of nicotine dependence. Thus, the presence of smoking-related cues would be required to induce craving and motivate substance use. In support of this argument, some studies found that lighter smokers exhibit greater attentional bias than heavier smokers (Bradley et al., **2003**; Hogarth et al., **2003**; Mogg et al., **2005**).

On the other hand, the argument could be made that heavier smokers should show greater attentional bias than lighter smokers since the conditioned response to smoking-related cues should be stronger due to repeated exposure. There is also evidence for this view as some studies have found that heavier smokers show greater attentional bias than lighter smokers (Chanon et al., **2010**; Vollstädt-Klein et al., **2011**; Zack et al., **2001**). Collectively, these studies show that smokers consistently display greater attentional bias towards smoking cues than non-smokers, but it is not clear whether lighter or heavier smokers show greater attentional bias.

To address this inconsistency, the current study focused on comparing attentional bias towards smoking cues in daily and non-daily smokers. While most studies use the visual probe task to measure attentional bias, their relatively small sample sizes and inconsistent research design features complicate drawing conclusions from the mixed findings. Therefore, we used a much larger sample size than previous studies and manipulated different features of the visual probe task.

The visual probe task infers attention through differences in response time (RT). Two images are presented and when they disappear, the participant is required to indicate the location of a small probe that replaces one of the images. Faster RTs to particular stimuli reflect selective attention (Field & Cox, **2008**), but as the location of attention is inferred through differences in RT after the stimuli disappear, the presentation time can be manipulated. Short Stimulus Onset Asynchronies (SOA) of 200ms or less measure involuntary attentional processes (Field & Cox, **2008**). Longer SOAs of 500ms or more target voluntary attention because there is enough time to make multiple fixations. Previous research used single longer SOAs of 500ms (Vollstädt-Klein et al., **2011**) and 2000ms (Hogarth et al., **2003**; Mogg et al., **2005**). None of the studies used a very short SOA to measure more involuntary attentional processes. Chanon et al. (**2010**) found that, in comparison to non-smokers, smokers' attentional bias was greater under 200ms conditions than a 550ms condition. To investigate the discrepancy in results between daily and non-daily smokers, this study used two SOAs of 200ms and 500ms.

A final consideration of our study was to evaluate and report the internal consistency of the visual probe task. There is growing awareness that the reliability of cognitive tasks should be taken seriously (Parsons et al., **2019**; Pennington et al., **2021**), but reliability has a different meaning depending on the context. For experimental measures to be reliable, we want to consistently observe effects between groups or conditions, but for correlational measures to be reliable, we want to consistently rank individuals (Hedge et al., **2018**). This means the attributes of experimental measures may not be

[1] Note, we refer to lighter and heavier smokers here as the studies used different definitions. In our study, we operationalize the groups as daily and non-daily smokers.

compatible with the requirements for reliable correlation research. As researchers often use the visual probe task as a measurement in cognitive bias modification procedures, it must be reliable to detect any changes across time. Previous attempts at evaluating the internal consistency of the visual probe task have been disappointing (Ataya et al., **2012**; Schmukle, **2005**; Waechter et al., **2014**). Therefore, we are following recommendations to habitually report the reliability of cognitive tasks (Parsons et al., **2019**), even when that is not the focus of the study.

The protocol and hypotheses for this project were pre-registered on the Open Science Framework (OSF; https:/ / osf.io/ t3xw8/). Given the relevance of smoking cues for non-daily smokers and the results from previously unpublished research, we hypothesized that non-daily smokers would show greater attentional bias than daily smokers. There was no *a priori* hypothesis for the effect of the SOA condition. This means that, though we expected non-daily smokers to show greater attentional bias than daily smokers, it was not clear what the difference in magnitude would be under different SOA conditions.

Method

Design

We used a 2 x 2 mixed design with one between-subjects independent variable (IV) of smoking group with two levels: daily and non-daily smokers. Participants responded to the question "Do you usually smoke cigarettes every day?". Non-daily smokers responded "No" and daily smokers responded "Yes". There was one within-subjects IV of the visual probe task SOA, which had two levels: 200ms and 500ms. The dependent variable (DV) was the attentional bias index (ms) calculated by subtracting the mean RT to smoking trials from the mean RT to non-smoking trials. Consequently, positive values would indicate greater attentional bias towards smoking cues.

Participants and Sample Size Calculation

We collected data online using Prolific where inclusion criteria consisted of: participants should have normal or corrected-normal vision; be between the ages of 18 to 60; and smoke at least one cigarette per week or four cigarettes per month.

We simulated a power analysis to justify the sample size. We set the smallest effect size of interest based on a previously unpublished study (Bartlett, **2020**) where the mean difference in attentional bias score between smoking groups was 6.13ms (95% confidence interval (*CI*) = [-5.27, 17.53]) for a 200ms SOA and 11.35ms (95% *CI* = [-4.51, 27.21]) for a 500ms SOA. However, we also consulted previous research due to the wide confidence intervals. The smallest known effects for a 200ms SOA were 5ms (Chanon et al., **2010**) and 11ms for a 500ms SOA (Bradley et al., **2003**). Our smallest effect sizes of interest were 5ms (200ms) and 10ms (500ms), and a conservative standard deviation of 20ms based on Vollstädt-Klein et al. (**2011**).

These values were used to conduct a simulated power analysis for a 2 x 2 mixed ANOVA using R (code available on the OSF; https://osf.io/ t3xw8/). We expected non-daily smokers to display greater attentional bias towards smoking images than daily smokers. We set the conditions of the power analysis as non-daily smokers having a 5ms (200ms) and 10ms (500ms) greater mean difference than daily smokers. For each condition, the values were sampled from a normal distribution with a standard deviation of 20ms. The sample size for each smoking group was increased from 10 (N = 20) to 150 (N = 300) in steps of 10, with each step repeating 10,000 times. The final sample size target was 60 per group (N = 120) as we reached 80% power (α = .05) between 50 and 60 participants per group.

Materials

Fagerström Test for Cigarette Dependence (FTCD)

The FTCD (Fagerström, **2012**; Heatherton et al., **1991**) was used as a self-report measure of nicotine dependence. The Cronbach's alpha estimate (bootstrapped using 10,000 iterations) in this sample was higher than in previous research, α = .74, 95% *CI* = [.67, .8].

Visual Probe Task

We used Gorilla (Anwyl-Irvine et al., **2019**) to

present the visual probe task online and the task is available on the open materials page to preview or clone (https://gorilla.sc/openmaterials/85021).

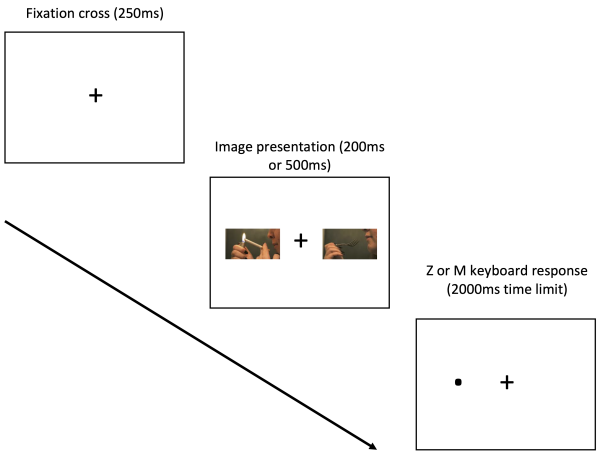


Figure 1 Diagram showing the trial procedure of the visual probe task. Each trial started with a fixation cross lasting 250ms. The fixation cross is then flanked by one of the stimulus pairs on the left and right. The stimuli remained on the screen for 200ms or 500ms depending on the SOA condition. The stimuli disappear and one image is replaced with a small dot. Participants had up to 2000ms to respond whether the dot was on the left or right. The next trial started with a new blank fixation cross.

Table 1 Mean (SD) values for participant characteristics and scale scores.

	Non-Daily Smokers	Daily Smokers
Age	28.68 (7.71)	31.84 (9.7)
% female	46.67%	26.42%
% white	93%	92%
FTCD	0.52 (1.31)	2.58 (2.17)
Cigarettes per day	2.38 (2.74)	8.59 (6.41)
Age started to smoke	18.51 (3.65)	17.93 (3.47)
Time since last cigarette (minutes)*	2880 (4590)	60 (633.75)

Note. *Due to large skew, these values represent the median and IQR.

Each trial started with a 250ms central fixation cross before two images were presented horizontally to the left and right. The content and duration of the two images was controlled by two variables: trial type and SOA. Trial type consisted of three conditions (neutral, smoking, or non-smoking) while SOA consisted of two conditions (200ms or 500ms). At picture offset, a small dot appeared in the location vacated by one of the images. The dot remained on the screen until the participant responded either left (Z key) or right (M key). After the participant responded the next trial began, with the screen containing only the fixation cross. The trial procedure is shown visually in Figure 1.

The trial type condition was based on 16 image pairs for neutral trials and 16 image pairs for smoking and non-smoking trials, meaning 32 unique image pairs in total. For neutral trials, the dot replaced either of the neutral image pairs. For smoking trials, the dot replaced a smoking image presented next to a matched non-smoking image. For non-smoking trials, the dot replaced a non-smoking image presented next to a smoking image.

We used 16 image pairs from the International Affective Picture System (Lang et al., **2008**) for the neutral trials. We developed a series of matching smoking and non-smoking images for the smoking and non-smoking trials (Bartlett, **2020**). The list of IAPS images is available on the OSF project (https://osf.io/fwud6/), and our smoking/non-smoking images are available on the Gorilla open materials page.

The trial order was randomized with each picture pair presented four times to cover each combination of image (left and right) and dot location (left and right). This combination determined the trial type condition, where a left smoking image, right non-smoking image, and left dot would produce a smoking trial. For each picture pair, this process was repeated twice for each SOA condition, producing 384 trials split into two blocks with 64 trials in each SOA and trial type condition.

Procedure

We provided participants with an information sheet, and they provided informed consent by ticking a box. This study was approved by the Ethical Approval board of the Faculty of Health and Life Sciences at Coventry University, United Kingdom (project reference number P88261). Participants completed a short questionnaire on their demographic information,

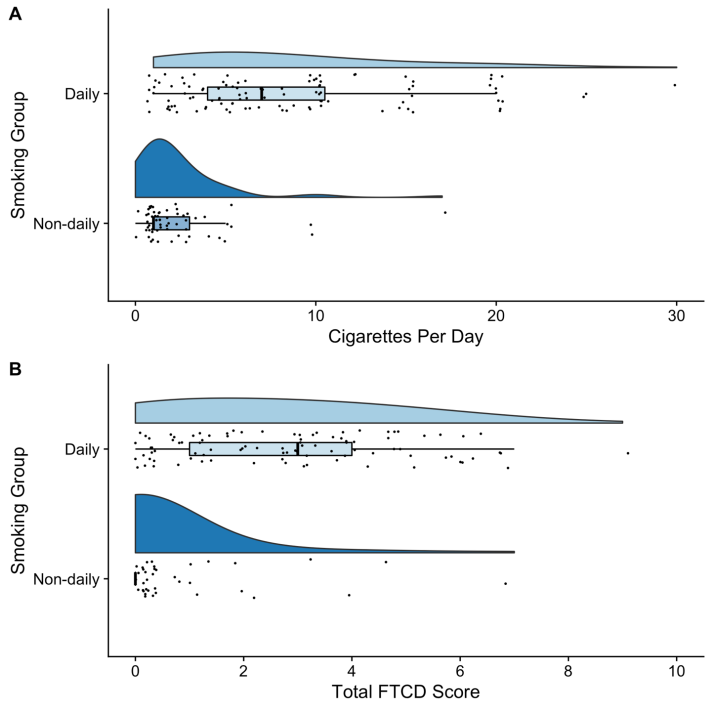


Figure 2 Two different measures of nicotine dependence: (A) number of cigarettes per day, and (B) FTCD score. The data are presented as raincloud plots (Allen et al., **2019**). The top element for each group represents the distribution of scores through the density. The bottom element presents the individual data points with a superimposed boxplot.

smoking habits, and the FTCD. The next page contained the visual probe task which began with a set of instructions asking the participant to complete the task in a quiet environment free of distractions. Participants completed 12 practice trials which provided feedback on their responses and overall accuracy. After the task, participants reported whether they experienced any technical issues, whether they used an ineligible device, and if they had completed the study before. Like Clifford and Jerit (**2014**), we asked participants if they had any distractions while they completed the study such as listening to music. Finally, participants read a debriefing sheet before they were redirected to Prolific. If the participants successfully reached the end of the study, they were paid £2.

Results

Participant Attrition and Demographics

In total, 218 people accessed the study, 205 of whom completed the experiment and received payment. The final sample was 166 after applying exclusion criteria: 60 non-daily and 106 daily smokers. Participants were excluded for having fewer than 50% of the possible trials (n = 4), experiencing technical issues (n = 16), reporting to smoke every day but not every week (n = 3), and not smoking in the past four weeks (n = 19). The total number of exclusions equals 42 as some participants met more than one criterion.

Table 1 displays the demographic information of the selected participants. Daily smokers smoked more cigarettes per day and had a higher FTCD score than non-daily smokers. Non-daily smokers clearly displayed infrequent smoking behavior as the median time since their last cigarette was 48 hours compared to only 1 hour for daily smokers. Figure 2 shows the distribution of FTCD scores and cigarettes per day.

Data Processing

The R code for all analyses is available on the OSF (https:/ / osf.io/ gm4jr/). We removed incorrect responses in addition to responses faster than 200ms as they represent preemptive responses. We considered outliers to be any response outside 2.5 times the median absolute deviation for each participant, SOA, and trial condition (Leys et al., **2013**). This meant we removed 9.72% of the total possible trials, with the median number of excluded trials for each participant being 23 (range 7 - 98).

For the confirmatory analyses, we focused on smoking/non-smoking image pairs and excluded the neutral pairs. Originally, we planned on conducting exploratory analyses to create orienting and disengagement indices (Salemink et al., **2007**) by subtracting the mean RT to neutral trials from smoking trials (orienting) or non-smoking trials (disengagement), but a coding error meant we did not have matching numbers of neutral trials in the 200ms and 500ms SOA conditions. Therefore, we focused on our confirmatory analyses and excluded neutral trials.

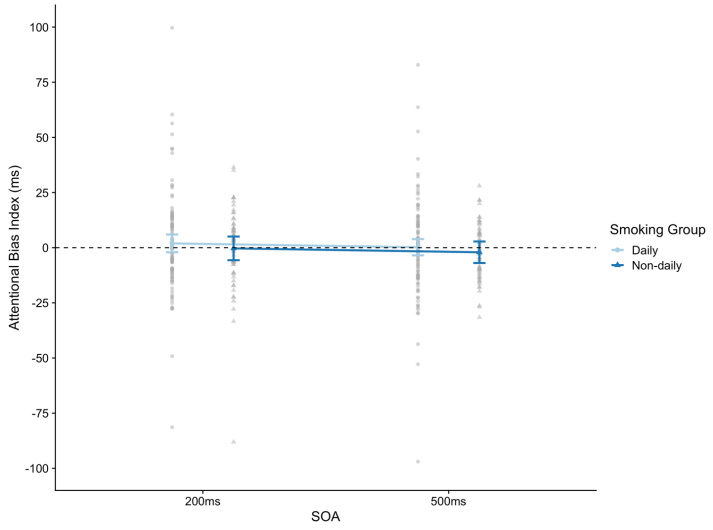


Figure 3 Interaction plot showing the mean attentional bias index for daily and non-daily smokers by SOA condition. The error bars represent the 95% CI around the mean. Positive values indicate greater attentional bias towards smoking cues. The grey points show the individual scores per condition.

After removing outliers, we calculated the mean RT to probes that replaced non-smoking images and the mean RT to probes that replaced smoking images. We then calculated the difference between these two values as our attentional bias index (non-smoking - smoking), where positive values mean faster average responses to smoking images. For each participant, this produced two values: one for the attentional bias index using a 200ms SOA and one for a 500ms SOA.

Confirmatory Analyses: Attentional Bias Towards Smoking Cues

The mean (*SD*) attentional bias index in the 200ms SOA condition was 1.95ms (22.31) for daily smokers and -0.30ms (18.57) for non-daily smokers. In the 500ms SOA condition, the mean bias index was 0.21ms (21.93) for daily smokers and -2.06ms (12.67) for non-daily smokers. This was in the opposite direction to our hypothesis as we expected non-daily smokers to display greater attentional bias towards smoking images than daily smokers. The detailed results are displayed in Figure 3.

We used a 2 x 2 mixed ANOVA with SOA as a within-subjects IV and smoking group as a between-subjects IV. The mean attentional bias index was the DV. There was not a significant effect of SOA (F (1, 164) = 0.58, p (1, 164) = 0.58, p = .448, $\hat{\eta}^2_G$ = .002) or smoking group (F (1, 164) = 0.58,p (1, 164) = 0.97, p = .325, $\hat{\eta}^2_G$ = .003). There was also no significant interaction between the two factors, F (1, 164) = 0.58, p (1, 164) = 0.01, p = .996, $\hat{\eta}^2_G$ < .001. These results do not support our prediction that non-daily smokers show greater attentional bias towards smoking images than daily smokers.

Exploratory Analyses: No Meaningful Difference in Attentional Bias

To demonstrate there was no meaningful difference between daily and non-daily smokers, we performed equivalence testing on the two comparisons of interest: the difference between daily and non-daily smokers at each SOA condition. One cannot directly provide evidence in favor of the null hypothesis using traditional null hypothesis significance testing. Equivalence testing applies two one-sided tests to user-defined boundaries representing effects considered too small to be practically or theoretically meaningful (Lakens et al., **2018**). If both tests are statistically significant, one can conclude that the observed effect size is statistically equivalent to zero based on the boundaries.

There are different approaches to setting the boundaries for your smallest effect size of interest. We used Cohen's d = ±0.41 based on the small telescopes method (Lakens et al., **2018**). The small telescopes method uses a sensitivity power analysis where you enter the sample size of a target study and calculate what effect size it would have 33% power to detect. In our case, we used two groups of 25 and 26 participants based on Vollstädt-Klein et al. (**2011**), which would have 33% power to detect an effect size of d = ±0.41 (α = .05). The small telescopes method is appropriate when previous research did not define their smallest effect size of interest, so it represents the effect size large enough to be detectable in the original study (Simonsohn, **2015**). Considering alternative choices for the effect size boundaries, our conclusions below hold when we use the larger effect size from our power analysis (10ms) but not when we use the smaller effect size (5ms). Because we are arguing differences

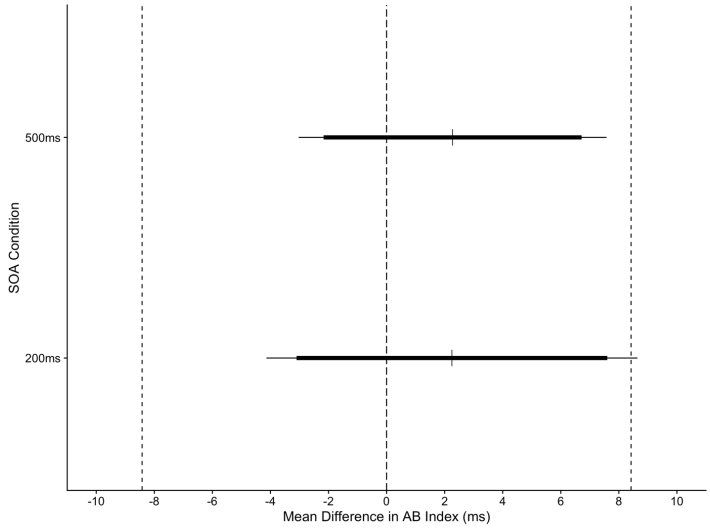


Figure 4 The thin vertical lines show the mean difference in attentional bias index between daily and non-daily smokers in each SOA condition. The thick horizontal black lines represent the 90% CI for the two one-sided test procedure. The thin horizontal black lines represent the 95% CI. The dashed vertical lines represent the equivalence boundaries in raw scores.

in attentional bias in daily and non-daily smokers may be smaller than reported in previous research, we focus on the results using the small telescopes method.

For the 200ms SOA condition, the two one-sided test procedure was significant, demonstrating that the difference in attentional bias towards smoking images between daily and non-daily smokers was statistically equivalent to zero, t (141.65) = -1.91, p = .029. Similarly, the 500ms SOA condition was statistically equivalent to zero, t (163.91) = -1.89, p = .03. The equivalence testing procedure is presented in Figure 4, showing that the 90% CI around the mean difference crosses zero, but does not cross the effect size boundaries of d = ±.41 (expressed here in raw units).

Exploratory Analyses: Including trial type as an additional IV

In our preregistration protocol, we focused on the attentional bias index as our outcome for confirmatory analyses, calculating it from the difference between smoking and neutral trials. While there were no meaningful differences between smoking groups, both peer-reviewers questioned whether participants first showed an attentional bias effect towards smoking images. Therefore, we performed exploratory analyses where we included trial type as an additional within-subjects IV instead of calculating the difference in RT between each condition.

We used a 2 x 2 x 2 mixed ANOVA using RT as our DV, trial type and SOA as within-subject IVs, and smoking group as a between-subjects IV. The only significant effect was SOA (F (1, 164) = 13.03, p < .001, $\hat{\eta}^2_G$ = .002), which in isolation is not theoretically meaningful to us. None of the other effects were statistically significant.

Although there were no significant effects including trial type, we quantified whether participants showed an attentional bias effect towards smoking images using the persons as effect sizes approach (Grice et al., **2020**). Instead of calculating a blanket mean difference between groups or conditions, one could quantify how many participants behaved consistent with theoretical predictions. In this context, we can ask how many participants showed faster RTs to smoking trials compared to non-smoking trials.

For each participant, we coded whether the difference in RT was negative (faster average responses to non-smoking images) or positive (faster average responses to smoking images), then calculated the percentage of participants showing a positive effect for each smoking group and SOA condition. Half (50%) of daily smokers in the 200ms and 52.83% in the 500ms SOA condition showed faster responses to smoking images. In contrast, 53.33% of non-daily smokers in the 200ms SOA condition showed faster responses to smoking images, while 43.33% responded faster to smoking images in the 500ms SOA condition, suggesting more participants responded faster to non-smoking images. We visualized these results in Figure 5 where each line represents a participant and the color shows whether they responded faster to smoking or non-smoking images for each SOA condition and smoking group. Collectively, these exploratory analyses suggest participants did not display the predicted attentional bias effect towards smoking images.

Exploratory Analyses: Visual Probe Task Reliability

Across the 16 smoking and non-smoking stimulus pairs, we calculated Cronbach's alpha for

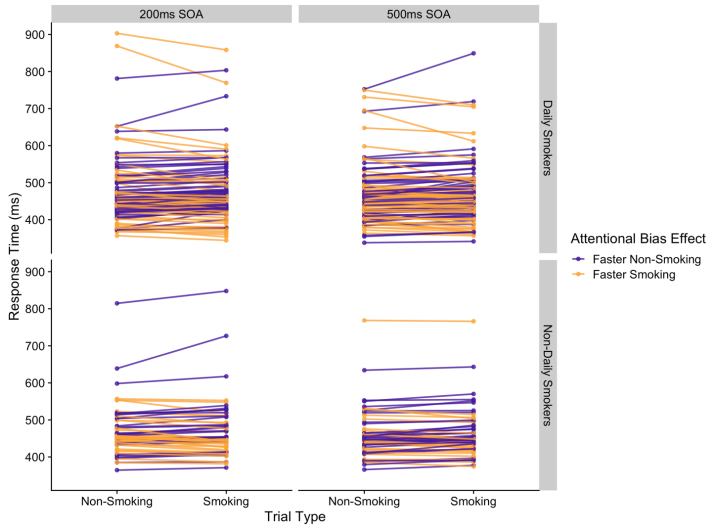


Figure 5 A dot plot visualizing whether each participant showed the predicted attentional bias effect towards smoking images. Each line represents one participant where their average RT to non-smoking and smoking images is connected. Positive slopes (purple lines) show participants who responded faster on average to non-smoking images while negative slopes (orange lines) show participants who responded faster on average to smoking images. Each panel represents the combination of smoking group and SOA condition.

the attentional bias index which was poor for both the 200ms (α = .29, 95% *CI* = [.00, .58]) and 500ms (α = .19, 95% *CI* = [.00, .42]) SOA conditions.

We reported internal consistency estimates for comparison with previous studies, but these assume the items or trials are presented in the same order (Parsons et al., **2019**). As cognitive tasks randomize trials, internal consistency may not be the best approach. An alternative is a permutation approach to calculating split-half reliability (Parsons, **2020**). This randomly splits the data set into two halves many times and calculates the average correlation between each half. Using 5000 iterations, poor reliability was also reflected in the split-half estimate (corrected using the Spearman-Brown formula) for the 200ms (r = .56, 95% *CI* = [.37, .7]) and 500ms (r = .47, 95% *CI* = [.27, .62]) SOA conditions.

Discussion

We hypothesized that non-daily smokers would display greater attentional bias towards smoking cues than daily smokers. Existing literature showed ambiguous results. Some studies found that non-daily smokers exhibited greater attentional bias (Bradley et al., **2003**; Hogarth et al., **2003**; Mogg et al., **2005**), whereas others found that daily smokers displayed greater attentional bias (Chanon et al., **2010**; Vollstädt-Klein et al., **2011**; Zack et al., **2001**). Using traditional methods that calculate an attentional bias index from average differences in RT, the current study found no significant differences, with equivalence testing showing that there was no meaningful difference in attentional bias in daily and non-daily smokers.

We may have found null results as previous research could have problems with inflated effect sizes due to low statistical power. The previous largest sample was 51 smokers in Vollstädt-Klein et al. (**2011**). Splitting these into 25 and 26 participants, a sensitivity power analysis indicates that this sample size would be sensitive to detect effect sizes of Cohen's d = 0.80 (α = .05, power = .80). Incidentally, Schäfer and Schwarz (**2019**) showed that the median Cohen's d in a random selection of 684 non-pre-registered articles was 0.80. In the long run, our study would have 99.80% power to detect an effect size of d = 0.80. Therefore, it is unlikely the effect size between daily and non-daily smokers is this large; if it was, we would have had enough power to detect it. Our study had the largest known sample size to investigate attentional bias with 60 non-daily smokers and 106 daily smokers. A sensitivity power analysis shows that this was sensitive to detect effect sizes of Cohen's d = 0.46. Our study was sensitive to detect an effect size of almost half the size of Vollstädt-Klein et al. (**2011**). Yet our results were statistically equivalent to zero, meaning there may not be a meaningful difference in attentional bias between smoking groups, at least in its current implementation where the effect is assumed to represent stable trait-like group differences.

Contemporary theories suggest attentional bias may not be a trait-like phenomenon that can produce stable differences between groups. Field et al. (**2016**) suggested that attentional bias varies depending on how substance cues are being evaluated. This theory suggests that rather than being a stable trait between

groups, attentional bias fluctuates with the incentive value of a cue, making within-group differences more important. Begh et al. (**2016**) found that laboratory measures like the visual probe task did not predict smoking behavior in the real-world. However, ecological momentary assessment of craving and awareness of smoking cues did predict smoking behavior. Therefore, the null results in our study may be a product of the fluctuating nature of attentional bias (Field et al., **2016**). In smaller samples, attentional bias could fluctuate one way or the other, but in larger samples (like our study) the differences could cancel out and converge to a mean difference around zero. Therefore, future research may benefit from investigating which factors affect the momentary evaluation of substance cues and the subsequent expression of attentional bias.

Using the visual probe task to measure factors that affect the momentary evaluation of substance cues may be problematic, though. There are vocal critics of the task due to its questionable level of internal consistency (Ataya et al., **2012**; Schmukle, **2005**; Waechter et al., **2014**). Our study also had suboptimal levels of internal consistency and split-half reliability. Researchers rarely report the reliability of cognitive tasks unless it is the focus of the article (Parsons et al., **2019**), which means it is difficult to assess how reliable the tasks were in previous smoking research. Experimental measures are designed to produce reliable differences between groups or condition, not consistently rank individuals (Hedge et al., **2018**). This means if researchers plan to use the visual probe task across multiple measurements - such as in cognitive bias modification or the evaluation of substance cues - its poor reliability is problematic. Future research should consider using eye-tracking as a direct measure of attentional bias as it produces larger effect sizes (Field et al., **2009**), has higher internal consistency (Price et al., **2015**), and higher criterion validity (Soleymani et al., **2020**).

Original Purpose

Daily and non-daily smokers have different habits and motives but both groups find it difficult to quit smoking long-term. As attentional bias may be associated with addictive behavior, we used the visual probe task to compare daily and non-daily smokers. We predicted that non-daily smokers would show greater attentional bias towards smoking images than daily smokers. If non-daily smokers showed greater attentional bias, it would help to explain why they find it difficult to quit smoking while showing fewer signs of nicotine dependence.

Limitations

Our sample may have been more diverse in age and education than typical undergraduate samples, but it still contained predominantly white participants. Non-daily smoking is more prevalent in ethnic minority groups (Fagan & Rigotti, **2009**; Levy et al., **2009**) and the health implications of smoking disproportionately affect non-white smokers (St.Helen et al., **2019**). Therefore, future research would benefit from recruiting a larger proportion of non-white smokers for the results to generalize beyond mostly white smokers.

The online nature of the study meant participants' smoking levels could not be verified objectively using measures like Carbon Monoxide (Wray et al., **2016**), but Ramo et al. (**2011**) demonstrated that smoking-related information collected online has good reliability and validity. Relatedly, as participants completed the study online, there was no control over their smoking behavior before and during the study. This led to idiosyncrasies as some smokers reported smoking while they were completing the study. Although this may represent a more naturalistic environment for the smokers, our study had less control over smokers' deprivation levels.

Conclusion

The purpose of our study was to investigate the conflict in attentional bias results between daily and non-daily smokers. We expected non-daily smokers to show greater attentional bias towards smoking images than daily smokers. Greater attentional bias in non-daily smokers would have helped to explain why they find it difficult to quit smoking while showing fewer signs of nicotine dependence. However, there were no significant effects and using equivalence testing, we found that there was no

meaningful difference in attentional bias between daily and non-daily smokers. The results can be interpreted in line with contemporary theories of attentional bias where there may not be stable trait-level differences between smoking groups in attentional bias. Future research should focus on investigating how attentional bias fluctuates over time using more reliable measures than the visual probe task.

Disclosures

CRediT Contributions

Conceptualization (JEB, RJ, NW); Methodology (JEB, RJ, NW); Formal analysis (JEB); Investigation (JEB); Data curation (JEB); Writing - original draft (JEB); Writing - Review & editing (JEB); Supervision (RJ, NW).

Data, code, and materials

The data and code to reproduce these analyses are available on the OSF (https://osf.io/am9hd/). The OSF project contains all necessary files to reproduce the analyses and figures. The visual probe task was created in Gorilla and the task can be found using the open materials page (https://gorilla.sc/openmaterials/85021)

R Package Acknowledgements

The results were created using R (version 4.1.3 R Core Team, **2020**) and the R-packages *afex* (version 1.0.1 Singmann et al., **2020**), *cowplot* (version 1.1.1 Wilke, **2019**), *dplyr* (version 1.0.10 Wickham & Henry, **2020**), *ggplot2* (version 3.3.5 Wickham, **2016**), *janitor* (version 2.1.0 Firke, **2019**), *papaja* (version 0.1.1 Aust & Barth, **2022**), *psych* (version 2.2.3 Revelle, **2019**), *pwr* (version 1.3.0 Champely, **2020**), *readr* (version 2.1.2 Wickham et al., **2018**), *shiny* (version 1.7.1 Chang et al., **2020**), *splithalf* (version 0.8.2 Parsons, **2020**), *stringr* (version 1.4.0 Wickham, **2019**), *tibble* (version 3.1.6 Müller & Wickham, **2020**), *tidyr* (version 1.2.0 Wickham & Henry, **2020**), *tinylabels* (version 0.2.3 Barth, **2022**), and *TOSTER* (version 0.4.0 Lakens, **2017**).

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., & Kievit, R. A. (2019). Raincloud plots: A multi-platform tool for robust data visualization. *Wellcome Open Research*, *4*, 63. https://doi.org/10.12688/wellcomeopenres.15191.1 (see p. 5).

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. https://doi.org/10.1101/438242 (see p. 3).

Ataya, A. F., Adams, S., Mullings, E., Cooper, R. M., Attwood, A. S., & Munafò, M. R. (2012). Internal reliability of measures of substance-related cognitive bias. *Drug and Alcohol Dependence*, *121*(1), 148–151. https://doi.org/10.1016/j.drugalcdep.2011.08.023 (see pp. 3, 9).

Aust, F., & Barth, M. (2022). *papaja: Prepare reproducible APA journal articles with R Markdown* [R package version 0.1.1]. https://github.com/crsh/papaja (see p. 10).

Barth, M. (2022). Tinylabels: Lightweight variable labels. https://cran.r-project.org/package=tinylabels (see p. 10).

Bartlett, J. E. (2020). *Daily and non-daily smokers: A profile of drive and cognitive control mechanisms* [PhD]. Coventry University. https://thesiscommons.org/h9gpe/ (see pp. 3, 4).

Baschnagel, J. S. (2013). Using mobile eye-tracking to assess attention to smoking cues in a naturalized environment. *Addictive Behaviors*, *38*(12), 2837–2840. https://doi.org/10.1016/j.addbeh.2013.08.005 (see p. 2).

Begh, R., Smith, M., Ferguson, S. G., Shiffman, S., Munafò, M. R., & Aveyard, P. (2016). Association between smoking-related attentional bias and craving measured in the clinic and in the natural environment. *Psychology of Addictive Behaviors*, *30*(8), 868–875. https://doi.org/10.1037/adb0000231 (see p. 9).

Bogdanovica, I., Godfrey, F., McNeill, A., & Britton, J. (2011). Smoking prevalence in the european union: A comparison of national and transnational prevalence survey methods and results. *Tobacco Control*, *20*(1), 1–9. https://doi.org/10.1136/tc.2010.036103 (see p. 1).

Bradley, B. P., Mogg, K., Wright, T., & Field, M. (2003). Attentional bias in drug dependence: Vigilance for cigarette-related cues in smokers. *Psychology of Addictive Behaviors*, *17*(1), 66–72. https://doi.org/10.1037/0893-164X.17.1.66 (see pp. 2, 3, 8).

Champely, S. (2020). Pwr: Basic functions for power analysis. https://CRAN.%20R-project.org/package=pwr (see p. 10).

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). Shiny: Web application framework for r. https://CRAN.%20R-project.org/package=shiny (see p. 10).

Chanon, V. W., Sours, C. R., & Boettiger, C. A. (2010). Attentional bias toward cigarette cues in active smokers. *Psychopharmacology*, *212*(3), 309–320. https://doi.org/10.1007/s00213-010-1953-1 (see pp. 2, 3, 8).

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? an experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(2), 120–131. https://doi.org/10.1017/xps.2014.5 (see p. 5).

Ehrman, R. N., Robbins, S. J., Bromwell, M. A., Lankford, M. E., Monterosso, J. R., & O'Brien, C. P. (2002). Comparing attentional bias to smoking cues in current smokers, former smokers, and non-smokers using a dot-probe task. *Drug and Alcohol Dependence*, *67*(2), 185–191. https://doi.org/10.1016/S0376-8716(02)00065-0 (see p. 2).

Fagan, P., & Rigotti, N. A. (2009). Light and intermittent smoking: The road less traveled. *Nicotine & Tobacco Research*, *11*(2), 107–110. https://doi.org/10.1093/ntr/ntn015 (see pp. 1, 9).

Fagerström, K. (2012). Determinants of tobacco use and renaming the ftnd to the fagerström test for cigarette dependence. *Nicotine & Tobacco Research*, *14*(1), 75–78. https://doi.org/10.1093/ntr/ntr137 (see p. 3).

Field, M., & Cox, W. M. (2008). Attentional bias in addictive behaviors: A review of its development, causes, and consequences. *Drug and Alcohol Dependence*, *97*(1-2), 1–20. https://doi.org/10.1016/j.drugalcdep.2008.03.030 (see pp. 1, 2).

Field, M., Munafò, M. R., & Franken, I. H. (2009). A meta-analytic investigation of the relationship between attentional bias and subjective craving in substance abuse. *Psychological Bulletin*, *135*(4), 589–607. https://doi.org/10.1037/a0015843 (see pp. 1, 9).

Field, M., Werthmann, J., Franken, I., Hofmann, W., Hogarth, L., & Roefs, A. (2016). The role of attentional bias in obesity and addiction. *Health Psychology*, *35*(8), 767–780. https://doi.org/10.1037/hea0000405 (see pp. 1, 8, 9).

Firke, S. (2019). Janitor: Simple tools for examining and cleaning dirty data. https://CRAN.%20R-project.org/package=janitor (see p. 10).

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as effect sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982 (see p. 7).

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerström, K.-O. (1991). The fagerström test for nicotine dependence: A revision of the fagerstrom tolerance questionnaire. *British Journal of Addiction*, *86*(9), 1119–1127. https://doi.org/10.1111/j.1360-0443.1991.tb01879.x (see p. 3).

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1 (see pp. 2, 9).

Hogarth, L. C., Mogg, K., Bradley, B. P., Duka, T., & Dickinson, A. (2003). Attentional orienting towards smoking-related stimuli. *Behavioural Pharmacology*, *14*(2), 153–160. https://doi.org/10.1097/00008877-200303000-00007 (see pp. 2, 8).

Kang, O.-S., Chang, D.-S., Jahng, G.-H., Kim, S.-Y., Kim, H., Kim, J.-W., Chung, S.-Y., Yang, S.-I., Park, H.-J., Lee, H., & Chae, Y. (2012). Individual differences in smoking-related cue reactivity in smokers: An eye-tracking and fmri study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *38*(2), 285–293. https://doi.org/10.1016/j.pnpbp.2012.04.013 (see p. 2).

Kotz, D., Fidler, J., & West, R. (2012). Very low rate and light smokers: Smoking patterns and cessation-related behaviour in england, 2006-11: Very low rate and light smokers. *Addiction*, *107*(5), 995–1002. https://doi.org/10.1111/j.1360-0443.2011.03739.x (see p. 1).

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. https://doi.org/10.1177/1948550617697177 (see p. 10).

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963 (see p. 6).

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (iaps): Affective ratings of pictures and instruction manual*. University of Florida. (See p. 4).

Levy, D. E., Biener, L., & Rigotti, N. A. (2009). The natural history of light smokers: A population-based cohort study. *Nicotine & Tobacco Research*,

11(2), 156–163. https://doi.org/10.1093/ntr/ntp011 (see p. 9).

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013 (see p. 5).

Mogg, K., Bradley, B. P., Field, M., & Houwer, J. (2003). Eye movements to smoking-related pictures in smokers: Relationship between attentional biases and implicit and explicit measures of stimulus valence. *Addiction*, *98*(6), 825–836. https://doi.org/10.1046/j.1360-0443.2003.00392.x (see p. 2).

Mogg, K., Field, M., & Bradley, B. P. (2005). Attentional and approach biases for smoking cues in smokers: An investigation of competing theoretical views of addiction. *Psychopharmacology*, *180*(2), 333–341. https://doi.org/10.1007/s00213-005-2158-x (see pp. 2, 8).

Müller, K., & Wickham, H. (2020). Tibble: Simple data frames. https://CRAN.%20R-project.org/package=tibble (see p. 10).

Parsons, S. (2020). Splithalf; robust estimates of split half reliability. https://doi.org/10.6084/m9.figshare.5559175.v5 (see pp. 8, 10).

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695 (see pp. 2, 3, 8, 9).

Pennington, C. R., Jones, A., Bartlett, J. E., Copeland, A., & Shaw, D. J. (2021). Raising the bar: Improving methodological rigour in cognitive alcohol research. *Addiction*, *116*(11), 3243–3251. https://doi.org/10.1111/add.15563 (see p. 2).

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., Dahl, R. E., & Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment*, *27*(2), 365–376. https://doi.org/10.1037/pas0000036 (see p. 9).

R Core Team. (2020). R: A language and environment for statistical computing. r foundation for statistical computing. https://www.R-project.org/ (see p. 10).

Ramo, D. E., Hall, S. M., & Prochaska, J. J. (2011). Reliability and validity of self-reported smoking in an anonymous online survey with young adults. *Health Psychology*, *30*(6), 693–701. https://doi.org/10.1037/a0023443 (see p. 9).

Revelle, W. (2019). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.%20R-project.org/package=psych (see p. 10).

Salemink, E., Hout, M. A. d., & Kindt, M. (2007). Selective attention and threat: Quick orienting versus slow disengagement and two versions of the dot probe task. *Behaviour Research and Therapy*, *45*(3), 607–615. https://doi.org/10.1016/j.brat.2006.04.004 (see p. 5).

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 1–13. https://doi.org/10.3389/fpsyg.2019.00813 (see p. 8).

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality*, *19*(7), 595–605. https://doi.org/10.1002/per.554 (see pp. 3, 9).

Shiffman, S. (2009). Light and intermittent smokers: Background and perspective. *Nicotine & Tobacco Research*, *11*(2), 122–125. https://doi.org/10.1093/ntr/ntn020 (see p. 1).

Shiffman, S., Dunbar, M. S., Li, X., Scholl, S. M., Tindle, H. A., Anderson, S. J., & Ferguson, S. G. (2014). Smoking patterns and stimulus control in intermittent and daily smokers. *PLoS ONE*, *9*(3), 1–14. https://doi.org/10.1371/journal.pone.0089911 (see p. 1).

Shiffman, S., Dunbar, M. S., Scholl, S. M., & Tindle, H. A. (2012). Smoking motives of daily and non-daily smokers: A profile analysis. *Drug and Alcohol Dependence*, *126*(3), 362–368. https://doi.org/10.1016/j.drugalcdep.2012.05.037 (see p. 1).

Shiffman, S., Tindle, H., Li, X., Scholl, S., Dunbar, M., & Mitchell-Miland, C. (2012). Characteristics and smoking patterns of intermittent smokers. *Experimental and Clinical Psychopharmacology*, *20*(4), 264–277. https://doi.org/10.1037/a0027546 (see p. 1).

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341 (see p. 6).

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). Afex: Analysis of factorial experiments. https://CRAN.%20R-project.org/package=afex (see p. 10).

Soleymani, A., Ivanov, Y., Mathot, S., & Jong, P. J. (2020). Free-viewing multi-stimulus eye tracking

task to index attention bias for alcohol versus soda cues: Satisfactory reliability and criterion validity. *Addictive Behaviors*, *100*, 106117. https://doi.org/10.1016/j.addbeh.2019.106117 (see p. 9).

St.Helen, G., Benowitz, N. L., Ahluwalia, J. S., Tyndale, R. F., Addo, N., Gregorich, S. E., Pérez-Stable, E. J., & Cox, L. S. (2019). Black light smokers: How nicotine intake and carcinogen exposure differ across various biobehavioral factors. *Journal of the National Medical Association*, *111*(5), 509–520. https://doi.org/10.1016/j.jnma.2019.04.004 (see p. 9).

Tindle, H. A., & Shiffman, S. (2011). Smoking cessation behavior among intermittent smokers versus daily smokers. *American Journal of Public Health*, *101*(7), 1–3. https://doi.org/10.2105/AJPH.2011.300186 (see p. 1).

Tong, E. K., Ong, M. K., Vittinghoff, E., & Pérez-Stable, E. J. (2006). Nondaily smokers should be asked and advised to quit. *American Journal of Preventive Medicine*, *30*(1), 23–30. https://doi.org/10.1016/j.amepre.2005.08.048 (see p. 1).

Vollstädt-Klein, S., Loeber, S., Winter, S., Leménager, T., Goltz, C. d., Dinter, C., Koopmann, A., Wied, C., Winterer, G., & Kiefer, F. (2011). Attention shift towards smoking cues relates to severity of dependence, smoking behavior and breath carbon monoxide. *European Addiction Research*, *17*(4), 217–224. https://doi.org/10.1159/000327775 (see pp. 2, 3, 6, 8).

Waechter, S., Nelson, A. L., Wright, C., Hyatt, A., & Oakman, J. (2014). Measuring attentional bias to threat: Reliability of dot probe and eye movement indices. *Cognitive Therapy and Research*, *38*(3), 313–333. https://doi.org/10.1007/s10608-013-9588-2 (see pp. 3, 9).

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org (see p. 10).

Wickham, H. (2019). Stringr: Simple, consistent wrappers for common string operations. https://CRAN.%20R-project.org/package=stringr (see p. 10).

Wickham, H., & Henry, L. (2020). Tidyr: Tidy messy data. https://CRAN.%20R-project.org/package=tidyr (see p. 10).

Wickham, H., Hester, J., & Francois, R. (2018). Readr: Read rectangular text data. https://CRAN.%20R-project.org/package=readr (see p. 10).

Wilke, C. O. (2019). Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'. https://CRAN.%20R-project.org/package=cowplot (see p. 10).

Wray, J. M., Gass, J. C., Miller, E. I., Wilkins, D. G., Rollins, D. E., & Tiffany, S. T. (2016). A comparative evaluation of self-report and biological measures of cigarette use in non-daily smokers. *Psychological Assessment*, *28*(9), 1043–1050. https://doi.org/10.1037/pas0000227 (see p. 9).

Zack, M., Belsito, L., Scher, R., Eissenberg, T., & Corrigall, W. A. (2001). Effects of abstinence and smoking on information processing in adolescent smokers. *Psychopharmacology*, *153*(2), 249–257. https://doi.org/10.1007/s002130000552 (see pp. 2, 8).

Gamified Inoculation Against Misinformation in India: A Randomized Control Trial

Trisha Harjani[1], Melisa-Sinem Basol[1], Jon Roozenbeek[1], Sander van der Linden[1]

Although the spread of misinformation is a pervasive and disruptive global problem, extant research is skewed towards "WEIRD" countries leaving questions about how to tackle misinformation in the developing world with different media and consumption patterns unanswered. We report the results of a game-based intervention against misinformation in India. The game is based on the mechanism of psychological inoculation; borrowed from the medical context, inoculation interventions aim to pre-emptively neutralize falsehoods and help audiences spot and resist misinformation strategies. Though the efficacy of these games has been repeatedly demonstrated in samples from Western countries, the present study conducted in north India (n = 757) did not replicate earlier findings. We found no significant impact of the intervention on the perceived reliability of messages containing misinformation, confidence judgments, and willingness to share information with others. Our experience presents a teachable moment for the unique challenges associated with complex cultural adaptations and field work in rural areas. These results have significant ramifications for designing misinformation interventions in developing countries where misinformation is largely spread via encrypted messaging applications such as WhatsApp. Our findings contribute to the small but growing body of work looking at how to adapt misinformation interventions to cross-cultural settings.

Keywords *misinformation, India, inoculation theory, pre-bunking, WhatsApp*

[1] University of Cambridge

Received
June 7, 2022
Accepted
October 10, 2022
Published
February 27th, 2023
Issued
December 18, 2023

Correspondence
University of Cambridge
th649@cam.ac.uk

The spread of misinformation online is widely documented as a threat to democracies worldwide (Lewandowsky et al., **2017**; van der Linden, Maibach, et al., **2017**). In India, the world's largest democracy, the sharing of misinformation online has been linked to mob violence, and even killings (Arun, **2019**; Sundar et al., **2021**; Vasudeva & Barkdull, **2020**). While social media platforms such as Facebook or Twitter can flag misinformed content or remove it from their platforms, mobile instant messenger services such as WhatsApp and Telegram are limited by their end-to-end encrypted nature (Banaji et al., **2019**). Private conversations or groups form a closed network where misinformation can freely circulate without monitoring and studies have shown that this takes place in India (Badrinathan, **2021**), as well as Burundi (Mumo, **2021**), Nigeria, Brazil, and Pakistan (Pasquetto et al., **2020**). Furthermore, a significant proportion of the misinformation shared in India continues to be shared and circulated on WhatsApp even after being falsified by professional, third-party fact checkers (Reis et al., **2020**). This trend has created a breeding ground for unverified, misleading, or false information, some of which originates from political parties (Chibber & Verma, **2018**). Despite WhatsApp's countermeasures, which include implementing digital literacy programs, placing restrictions on forwarding, and broadcasting awareness-raising adverts, misinformation on the platform is persistent and has been exacerbated by COVID-19 (Al-Zaman, **2021**; Ferrara, **2020**). Given the limitations of implementing algorithmic solutions on private messaging platforms (Reis et al., **2020**), user-level solutions are an increasingly important avenue of research.

The overwhelming majority of individual-

Take-home Message

This study found that gamified inoculation interventions, which have worked well in Western countries, did not confer psychological resistance against misinformation to participants in India. This null result (possibly due to lower digital literacy rates) calls for further investigation into bottom-up interventions tackling misinformation on messaging platforms in developing countries.

level misinformation interventions have been tested on populations from developed, Western countries. This is indeed a feature of behavioral science in general where non-WEIRD (western, educated, industrialized, rich and democratic) samples are underrepresented (Henrich et al., **2010**; Rad et al., **2018**). There are several factors that could impede the generalizability of findings to India specifically. Since 2017, year-on-year internet penetration in India has grown by 13% in rural areas compared to 4% in urban neighborhoods (Bhattacharjee et al., **2021**). While misinformation can be spread by both urban and rural residents, the latter are likely to access the internet via 2G networks with limited resources for fact checking and a tendency to distribute WhatsApp messages with low reflexivity, as a mode of group participation or strategy to avoid feelings of exclusion (Banaji et al., **2019**). Given the collectivist culture in India (Kapoor et al., **2003**; Verma & Triandis, **2020**), even amongst youth samples (Rao et al., **2013**), the importance of group identities is heightened. Political parties frequently capitalize on these divisions, often along religious lines (Vaishnav et al., **2019**). Furthermore, the institutionalization of misinformation dissemination by political parties in India, whereby 'IT cells' troll and spread automated content, is not uncommon (Campbell-Smith & Bradshaw, **2019**) as part of their campaigning strategy (Banaji et al., **2019**).

To counter the spread of misinformation, several strategies have been researched at the individual level, the most well-known of which include fact-checking and "debunking" or correcting false information after exposure (Ecker et al., **2022**; van der Linden, **2022**; Walter & Murphy, **2018**). Studies examining the efficacy of such corrective measures have revealed mixed results. Although some have found that fact-checking can improve accuracy assessments (Clayton et al., **2020**; Porter & Wood, **2021**; Walter & Murphy, **2018**), there are several drawbacks to correcting misinformation post-exposure. One major issue concerns the continued influence of misinformation or the tendency for people to continue making inferences based on misinformation. They do so even when they acknowledge a correction (Ecker et al., **2022**; Lewandowsky et al., **2012**), which limits the correction's potential effectiveness. This is further compounded by the finding that (a) not all audiences are receptive to fact-checks (Walter et al., **2020**), (b) repeated exposure to misinformation can increase its perceived accuracy (Pennycook et al., **2018**; Swire et al., **2017**), and (c) that corrections do not scale, meaning they rarely reach the same number people as the initial misinformation (Roozenbeek & van der Linden, **2019**; van der Linden, **2022**). Lastly, corrective strategies are also difficult to implement on private messaging platforms given the invisibility of information flow in this sphere (Reis et al., **2020**).

Accordingly, studies which have evaluated fact-checking and literacy interventions in developing countries have revealed inconclusive results. For example, Guess et al. (**2020**) tested the effect of providing U. S. and Indian participants with tips on how to spot misinformation. They found a positive impact on people's ability to detect false information in the U. S. and in a highly educated online Indian sample, but not in a face-to-face sample obtained in rural Northern India. Similarly, Badrinathan (**2021**) tested the impact of an intensive one-hour in-person media literacy training during the 2019 national election and found no significant beneficial effects.

One study tested the impact of a debunking intervention via WhatsApp broadcast messaging in Zimbabwe, another country with high WhatsApp usage, finding that participants had increased knowledge about COVID-19 (Bowles et al., **2020**). Pasquetto et al. (**2020**) further found that, while corrections on encrypted group chats reduced belief in misinformation in India and Pakistan, WhatsApp users report corrections as unusual and socially awkward. Given the known challenges surrounding debunking and fact-checking, a promising ef-

fort against misinformation has been to pre-emptively debunk (or *prebunk*) falsehoods to allow individuals to acquire skills to detect and resist misinformation in the future (Lewandowsky & van der Linden, **2021**). This approach is based on the theory of psychological inoculation (McGuire, **1961**).

Theoretical Background: Prebunking and Inoculation Theory

Inoculation theory was originally developed in the 1960s and is based on the biological process of immunization (McGuire, **1961**, **1964**): just as exposure to a weakened dose of a pathogen can confer immunity against future infection(s), pre-emptively exposing people to weakened doses of misinformation—along with strong refutations—can cultivate cognitive immunity to future manipulation attempts. Inoculation theory has two key components. Firstly, the inoculation must have a forewarning to evoke threat or the motivation for people to defend themselves from a potential attack on their attitudes (Compton, **2012**). Being aware of one's vulnerability to manipulation is important for kick-starting resistance to persuasion (Sagarin et al., **2002**). Secondly, much like the injection of a weakened dose of a virus can build immunity through the production of antibodies, exposure to a weakened version of a persuasive argument along with a counterargument can inspire lowered vulnerability to misleading persuasion attempts (McGuire, **1961**). A meta-analysis of inoculation theory has found that it is effective at building resistance against persuasion across issues (Banas & Rains, **2010**).

In more recent years, the theory has informed the design of inoculation interventions aiming to endow attitudinal resistance against online misinformation specifically (for in-depth reviews see Compton et al., **2021**; Lewandowsky & van der Linden, **2021**; Roozenbeek & van der Linden, **2018**; van der Linden, **2022**). Some recent applications of inoculation theory include even potentially polarizing topics such as climate change (van der Linden, Leiserowitz, et al., **2017**), conspiracy theories (Banas & Miller, **2013**), or vaccinations (Jolley & Douglas, **2017**). However, all these studies aimed to inoculate people against misinformation about a specific issue. As such, they do not necessarily imply that the inoculation would be effective as a "broad-spectrum vaccine" against misinformation (Roozenbeek & van der Linden, **2018**). This prompted a shift away from narrow-spectrum inoculations to those that incorporate persuasion techniques common to misinformation more generally (Cook et al., **2017**; Roozenbeek & van der Linden, **2019**). In other words, familiarity with a weakened dose of the underlying *techniques* that are used to spread misinformation could impart an increased cognitive ability to detect manipulative information that makes use of such misinformation tactics. These tactics include emotionally manipulative language, group polarization, conspiratorial reasoning, trolling, and impersonations of fake experts, politicians, and celebrities (Roozenbeek & van der Linden, **2019**).

This strategy has demonstrated fairly consistent success (Basol et al., **2020**; Cook et al., **2017**; Roozenbeek & van der Linden, **2019**) including long-term efficacy, provided inoculated individuals are given short reminders or "booster shots" of the lessons learned (Maertens et al., **2021**). Yet, no study to date has tested the effect of inoculation interventions on the Indian population and inoculation researchers have noted a lack of generalizability of inoculation scholarship to non-WEIRD populations (Bonetto et al., **2018**), demanding interventions be adapted and evaluated.

Recent Applications: Inoculation Games

Recent applications of inoculation theory also depart from the traditional method of providing participants with ready-made counterarguments (so-called "passive inoculation") and instead use an "active" form of inoculation whereby participants themselves play an active role in generating resistance to manipulation (Roozenbeek & van der Linden, **2018**). Gamified interventions have proven to be a fruitful vehicle for active inoculation. One example of such an inoculation intervention is the online game *Bad News* (www.getbadnews.com): in this game, players find themselves in an artificial social media environment designed to mimic the features of widely used online platforms (Basol et al., **2020**; Maertens et al., **2021**; Roozenbeek & van der Linden, **2019**; Roozenbeek et al., **2021**). Across six levels, players

are warned about the dangers of fake news, and they develop an understanding of several widely used misinformation techniques through exposure to weakened dose of these tactics alongside ways to spot them. Evidence for the relative benefits of "active" inoculation is emerging (Basol et al., **2021**), particularly because it may strengthen associative memory networks, contributing towards higher resistance to persuasion (Pfau et al., **2005**).

However, the *Bad News* game, as well as two others (*Harmony Square* Roozenbeek & van der Linden, **2020**) and (*Go Viral!* Basol et al., **2021**), all focus on misinformation on public social media platforms (such as Facebook and Twitter). This reduces the potential applicability of these games in countries where direct messaging apps are a more common means of communication than public social media platforms. To address this problem, we engaged in a novel real-world collaboration with WhatsApp, Inc (Meta platforms) and developed a new game that inoculates people against misinformation on direct messaging apps, called *Join this Group* (link to English version; https://whatsapp.aboutbadnews.com). The Hindi-version of the game was tested in this study (further details in the method section). Its purpose is to inoculate participants against four manipulation techniques commonly present in misinformation on direct messaging apps. Specifically, these techniques are the impersonation of a fake expert (Goga et al., **2015**; Jung, **2011**; Reznik, **2013**), use of emotional language to frame content (Gross & Ambrosio, **2004**; Konijn, **2012**; Zollo et al., **2015**), polarization of narratives to create hostility towards the opposition (Groenendyk, **2018**; S. Iyengar & Krupenkin, **2018**), and the escalation of an issue such that misinformation triggers offline acts of aggression (BBC Monitoring, **2021**; Robb, **2021**).

The Present Research

This paper seeks to address two gaps in the literature on misinformation interventions. We first aim to understand whether inoculation against misinformation can improve people's ability to spot misinformation that is commonly shared in a private messaging context (such as on WhatsApp). Second, our sample is from India, an understudied population where the spread of misinformation via private messaging platforms has been linked to violence (McLaughlin, **2018**). We ran a field experiment in India testing the efficacy of the inoculation game, *Join this Group*.

This paper therefore makes two unique advancements to the literature. This study is the first to test an inoculation intervention against misinformation shared in the context of private messaging. This domain of information exchange is markedly different to public platforms such that the burden of identifying, addressing, and correcting misinformation falls on the user(s) (Pasquetto et al., **2020**). Moreover, we test the effectiveness of these modified interventions in India (n = 757), the largest market for WhatsApp globally (Findlay, **2019**). Both studies were approved by the Cambridge Psychology Research Ethics Committee (REC-2018-19/19). [Data and scripts are deposited on the Open Science Framework: https://osf.io/abjrg].

Method

We conducted a 2 (treatment – control) x 2 (pre – post) mixed-between randomized control trial on a sample collected from 8 North Indian states (Bihar, Chhattisgarh, Haryana, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh, and National Capital Territory (Delhi)). Participants were recruited as part of media literacy workshops administered to 1283 individuals. The experiment was conducted door-to-door, in person, with the assistance of iPads and smartphones through which participants could access the online intervention. After providing informed consent, participants were asked to indicate their frequency of WhatsApp usage in the last twelve months on a 5-point scale, ranging from "*Never*" to "*More than once a day*". Participants were then shown 16 screenshots of WhatsApp conversations in a randomized order (see Figure 1) and, following Roozenbeek et al. (**2021**), were asked to make three assessments: how reliable they found the post (1), how confident they are in their reliability assessment (2) and how likely they would be to share the message (3). All three assessments were rated on a 1-7 Likert scale (1 being "*Not at all*", 4 being "*Neutral*", and 7 being "*Very much*"). Of the 16 images, four were screenshots of authentic WhatsApp conversations, of which two

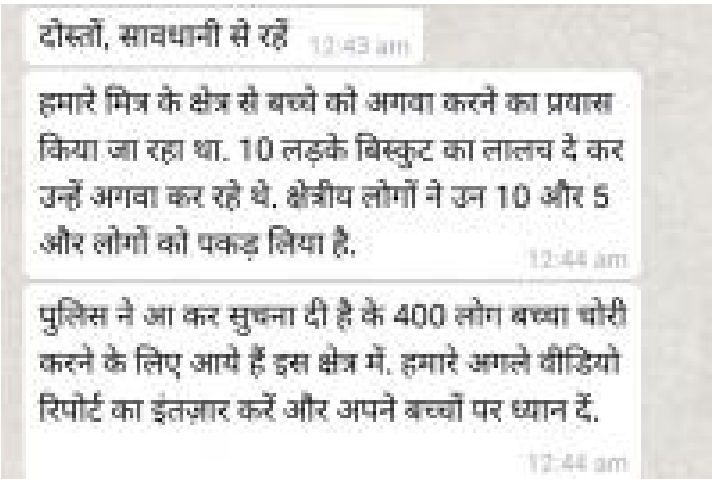


Figure 1 WhatsApp messages containing emotional misinformation messaging. This image is an example of one used in the experimental pre-test and post-test measure. The screenshot reads: "Friends, be careful", "Attempts are being made to kidnap a child from our friend's area. 10 boys were kidnapping him with the promise of biscuits. People in the area have caught those 10 and 5 more people", "The police has announced that 400 people had come to steal the child in this area. Wait for our next video that will report this and watch over your children carefully."

were fake news and two contained accurate information. The remaining 12 were screenshots containing misinformation designed to demonstrate four manipulation techniques (fake expert, emotion, polarization, and escalation). The four real (non-misinformation) items were sourced from fact-checking websites and the manipulative items were created by one of the authors and validated by two other authors, to ensure that the conversations make appropriate use of a misinformation technique. Figure 1 demonstrates an example of eliciting fear using emotional language in misinformation messaging.

Participants were then randomly assigned to play either *Join this Group* (treatment) or *Tetris* (control), consistent with previous gamified inoculation experiments (Basol et al., **2020**; Roozenbeek & van der Linden, **2020**). Gameplay for *Join this Group* was approximately 15 minutes while *Tetris* participants had to play for a minimum of nine minutes before proceeding. Participants who played *Join this Group* were required to input a password to validate their completion. Following the game, as part of the post-test measure, all participants were asked to assess the same 16 WhatsApp conversations again and answer some demographic questions, including district, state, gender, education level, age group, how frequently they check the news, how frequently they use social media platforms, their interest in politics, their political ideology, and attitudes scales assessing left to right and libertarian to authoritarian views (Park et al., **2013**). Participants were also asked to provide their first thoughts upon hearing the term "fake news."

Treatment Game: Join this Group

We created a Hindi translation of the *Join this Group* game in collaboration with a Delhi-based non-profit, the Digital Empowerment Foundation (DEF). One major challenge that arose during field implementation is that our novel inoculation approach did not fit conceptually into DEF's media literacy strategy. As a condition of administering the intervention in rural India, DEF therefore required that we adapt the intervention to be more in line with their own media literacy strategy. As a result, the key difference between the English and Hindi versions of the *Join this Group* game is that players take on more of a traditional fact-checking role by posing as an undercover detective fighting misinformation online. This is in stark contrast to active inoculation games such as *Bad News*, *GoViral!*, and *Harmony Square.* In these games, participants generally take on the role of a misinformation *spreader* because this perspective-taking exercise helps elicit "motivational threat" or the motivation to defend oneself against misinformation, a key component of inoculation theory (Basol et al., **2021**). However, DEF advised that such a perspective was not in line with their traditional media literacy training and may be confusing for their target audience in India, who generally have low digital literacy. Accordingly, we created a new version of the game where the player steps into the shoes of a fake news "detective."

In the Hindi version, players are introduced to the game with a messaging-interface screen reading "Hello detective! We need you." The game explains that a group called "Big News" is spreading propaganda on WhatsApp in the fictional nation of "Santhala." The game then explains that understanding the techniques

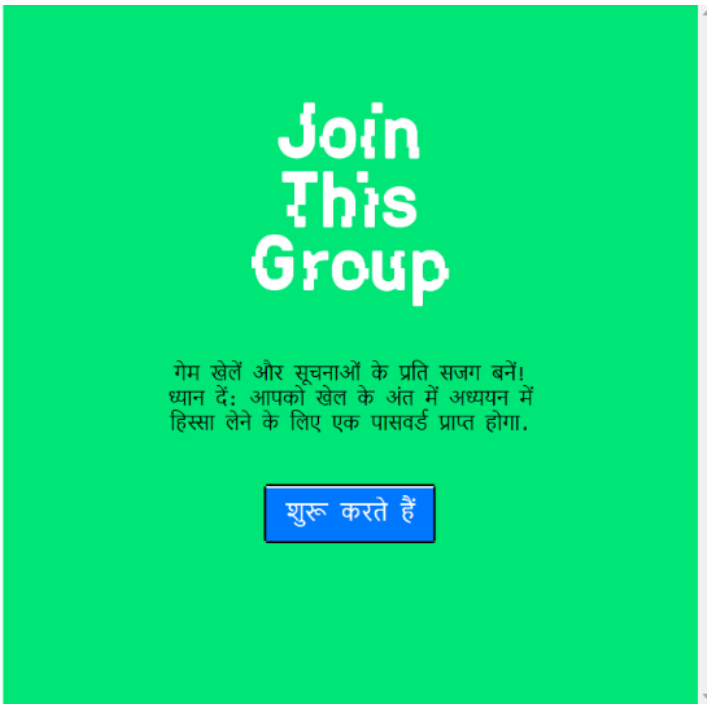


Figure 2 Landing page of the game. The text reads "Play the game and watch out for notifications! Attention: You will receive a password at the end of the game. In order to take part in the study, you'll need to input this password." Blue button reads "Let's start."

of the "Big News" group will require going undercover since messages are encrypted and untraceable. Figures 2 and 3 below display in-game screenshots. See Figures S4-S8[1] for more screenshots.

Players go through four levels, each one teaching and testing the application of techniques present in misinformation (fake experts, emotional language, polarization, escalation). See Table 1 for an overview of the four levels. In the first level, players are shown how sharing messages in a group unannounced can result in being reported, an issue that can be overcome by impersonating a fake expert to boost credibility of spurious claims. Players are then able to go undercover by spreading rumors such as "Mangoes cause cancer" using their fake pseudonym (See Figure 3). Such impersonations are pervasive throughout social media (Adewole et al., **2017**; Goga et al., **2015**; Jung, **2011**; Reznik, **2013**). The second level shows players how the use of emotionally charged language can create an atmosphere of chaos especially when combined with a visual prompt. Emotional framing and language have been shown to increase salience, social media engagement (Rathje et al., **2021**), grab attention (Konijn, **2012**), and evoke emotional reactions (Gross & Ambrosio, **2004**). The third level continues in context where players now need to apply their detective skills to prevent election manipulation. They are shown how repeated false messaging that uses partisan misinformation can vilify and antagonize the opposition (such as a political party), exaggerate the perceived distance between identities, sow doubt and increase support for a particular group (Groenendyk, **2018**; S. Iyengar & Krupenkin, **2018**; Melki & Pickering, **2014**). Finally, in the fourth level players are told that they need to report the partisan misinformation being shared. This results in the suspicion of a disloyal supporter in the political party's WhatsApp group and motivates a targeted offline attack on the mole, which intensifies into protests and riots. Throughout this level, the game explains how online encouragement can escalate into offline aggression (BBC Monitoring, **2021**; Robb, **2021**).

At the end of each level, players are given a summary of the techniques they have been inoculated against. Points and sanctions are also counted throughout; if players send a message that does not reflect use of the techniques learned, they are penalized. Conversely, exposing propaganda as an undercover detective increases points. In all scenarios, players also see WhatsApp group members' reactions to the misinformation. Overall, the game aims to demonstrate how fabricated content can evoke not only belief in misinformation but also create an atmosphere of fear, polarization, and elicit violent offline behavior.

The study was thus designed to test the efficacy of *Join this Group*, measured by three forms of assessment. We therefore hypothesized that:

H_1 Treatment group participants find manipulative WhatsApp messages significantly less reliable post-gameplay compared to the control group.

H_2 Treatment group participants are significantly more confident at assessing the reliabil-

[1] All figures and tables starting with S are to be found in the supplementary materials.

Figure 3 The first two messages after starting the game.
The top message reads "Hello Detective! We need you." The bottom message reads "Our great country Santhala needs you. A group called 'Big News' is spreading propaganda on a very large scale" (left).
In-game screenshot from the first level. The top message reads "Well done! Find the profile of a person who is a fake doctor." The bottom message reads "Dr. Saurav Agrawal" (right).

ity of manipulative WhatsApp messages compared to the control group.

H_3 Treatment group participants are significantly less likely to want to forward manipulative WhatsApp messages to others compared to the control group.

Sample

After providing informed consent, we collected n = 1283 observations, of which, n = 757 were complete responses. Participants did not always complete the full survey; we saw some drop-off after the intervention as many participants did not complete the post-test. To understand if the data was missing at random (MAR), we ran further analyses using the pre-test scores, condition allocation and WhatsApp usage data to assess missingness (see the supplementary materials for full details). We were not able to study the demographic predictors of the incomplete data because this was collected at the end of the study. The analysis finds that the data was not missing at random and that a higher baseline confidence in assessing the reliability of manipulative items decreased the odds of missingness (OR = 0.030, [95%CI; 0.002,0.431]) and being assigned to the treatment group increased the odds of missingness (OR = 2.171, [95%CI; 1.589, 2.967]). Please see Table S1 for full results.

During the data quality check, we further observed data in which participants just provided the same scale point consistently throughout the pre-test, post-test, or both (e.g., "4"). We therefore removed any responses which had repeated answer patterns[2] throughout the entire section (pre-test or post-test), resulting in a final sample size of n = 725. Of the final sample, 55% identified as female, 40% as male and 5% as other. 49% reported being 18-24 years old. 42% reported having obtained at least a bachelor's degree. The sample was also heavily left leaning, (M = 2.14, SD = 0.78). Finally, 65% of participants came from the state of Madhya Pradesh (17% from Rajasthan, 6% from Chhattisgarh, 5% from Uttar Pradesh, 4% from Jharkhand, 3% from Bihar). See Table S2 for a full breakdown of the sample.

Results

All data cleaning and analysis was conducted using RStudio, scripts are available via the Open Science Framework: https://osf.io/abjrg. For the main analyses, the following packages were used: stats (for ANCOVA), TOSTER (for tests of statistical equivalence) and BayesFactor (for Bayesian t-tests).

We conducted a one-way ANCOVA to test **H_1**,

[2]Analysis including the excluded 32 responses was also run and these did not affect the results.

Table 1 A summary of the game from the player's perspective at each of the four levels.

Level	*Manipulation Technique*	*Description*
1	Fake Expert	As undercover detectives, players join a WhatsApp group called "Breaking News" in the town of "Santhala." They share a fake message but are kicked out of the group, upon which they are encouraged to use a fake expert to gain credibility and witness how this impersonation can garner belief.
2	Emotional Language	Players are told that certain users in the group "Big News" are picking fights. As an undercover detective, they are tasked with spreading content to contribute to the chaos. The game then prompts players to share a fear or anger inducing message. This level shows players how, especially when paired with an image, emotional language can manipulate opinions and exacerbate chaos in the group.
3	Polarization	At this stage, Santhala is facing an election that the group "Breaking News" is attempting to manipulate. Players are told they must go undercover in one of the political candidate groups to spread polarizing information (e.g., damaging information about the opposition). The game shows how this cycle causes wider rifts between supporters.
4	Escalation	Continuing in context, the opposition group reports the polarizing fake news shared earlier to the media. The player is shown how members of the group try to identify the 'mole' which escalates into an offline attack on the suspected individual. Although WhatsApp now bans this political group, players are shown how they simply create another one with new phone numbers.

examining whether post-test reliability scores of manipulative items were significantly different between conditions, controlling for pre-test scores. We found no significant difference in reliability assessments between treatment and condition groups: $F(1,722) = 0.00$, $p = 0.97$. This relationship held for the subcategories of the fake items; fake expert: $F(1,722) = 0.21$, $p = 0.65$; emotion: $F(1,722) = 0.21$, $p = 0.65$; polarization: $F(1,722) = 0.35$, $p = 0.55$; and escalation: $F(1,722) = 0.03$, $p = 0.85$. To test whether the non-significant results imply null effects or equivalence to zero (Lakens et al., **2018**), we conducted an equivalence test using two one-sided tests (TOST) on the post-gameplay outcomes (TOSTs).[3] We could not confirm statistical equivalence to zero for the average reliability score $t(721.68) = -1.44$, $p = 0.07$. However, a Bayesian paired samples *t*-test for the averaged reliability score of misinformation items gives a Bayes factor of $BF_{10} = 0.25$ (error % = 0.00)**,** indicating support for the null hypothesis of $\mathbf{H_1}$ (Dienes, **2014**).

To test $\mathbf{H_2}$**,** we followed the same analysis: we conducted a one-way ANCOVA on the average post-test confidence in reliability judgment scores, controlling for the baseline. We find no significant difference between groups: $F(1,722) = 1.79$, $p = 0.18$ or for the subcategories; fake expert: $F(1, 722) = 1.56$, $p = 0.21$; emotion: $F(1,722) = 1.05$, $p = 0.31$; polarization: $F(1,722) = 1.18$, $p = 0.28$; escalation: $F(1,722) = 1.17$, $p = 0.28$. A TOST equivalence test confirmed equivalence to zero for the average post-test confidence scores (in assessing the reliability of misinformation items), $t(721.43) = -2.34$, $p = 0.01$. A Bayesian *t*-test provided strong evidence for the null hypothesis of $\mathbf{H_2}$, with a Bayes factor of $BF_{10} = 0.04$ (error % = 0.00).

To test $\mathbf{H_3}$, or whether there was a difference in post-test scores of intended willingness to share misinformation, another one-way ANCOVA was conducted on the average post-test scores, controlling for the baseline. Results were non-significant $F(1,722) = 1.46$, $p = 0.23$ including on the subcategories; fake expert: $F(1,722) = 1.94$, $p = 0.16$; emotion: $F(1,722) = 0.29$, $p = 0.59$; polarization: $F(1,722) = 2.75$, $p = 0.10$; and escalation: $F(1,722) = 2.77$, $p = 0.10$. A TOST analysis on the post-test likelihood to

[3]The smallest effect size of interest (SESOI) was set to d = ±0.25 based on the smallest observed effect size found in published experiments that use gamified inoculation interventions (Roozenbeek & van der Linden, **2019**).

share misinformation items scores could not confirm statistical equivalence to zero $t(719.73) = -0.64$, $p = 0.26$. However, a Bayesian t-test suggested strong support for the null hypothesis of $\mathbf{H_3}$ with a Bayes factor of $BF_{10} = 0.07$ (error % = 0.00). See Table S6 for Bayesian t-tests. Figure 4 shows the distribution of mean scores (reliability, confidence and sharing) for all misinformation items. Similarly, Figure 5 displays the distribution of mean reliability scores broken down by technique.

Though not hypothesized, to test whether the intervention increased skepticism towards factual messages, we also conducted a one-way ANCOVA to test for significant differences in post-gameplay scores for real news items, controlling for baseline scores. Specifically, ratings of reliability: $F(1,722) = 0.09$, $p = 0.76$; confidence in judgments: $F(1,722) = 1.10$, $p = 0.30$; and likelihood to share: $F(1,722) = 1.39$, $p = 0.24$ were not significantly different across treatment and control groups. Similarly, we tested whether the intervention improved participants assessments of the two genuine screenshots capturing fake news sharing on WhatsApp. Using one-way ANCOVAs we found no significant differences in ratings of reliability: $F(1, 712) = 0.99$, $p = 0.32$; confidence: $F(1, 711) = 1.68$, $p = 0.20$; or likelihood to share: $F(1,702) = 0.12$, $p = 0.73$.

We ran linear regressions to check for covariate effects on the differences in pre-post measures of reliability, confidence, and sharing. We only find that higher frequency of checking the news significantly predicts a larger difference between pre and post confidence scores of misinformation items ($p = 0.03$). See Tables S33-S35 for the full results.

Discussion

Through this study we demonstrate that there was no significant effect of playing *Join this Group* on the veracity evaluations of both real and misinformation items in our sample of North Indians. This is in contrast with previous results that have found promising results using gamified inoculation in Western populations (including versions translated to German, Greek, French, Polish, and Swedish (Basol et al., **2021**; Roozenbeek & van der Linden, **2020**). Direct replications of the *Bad News* game online have also shown positive effects on urban populations in India (A. Iyengar et al., **2022**) and importantly, randomized trial data[4] from a representative sample of the UK population using the English version of *Join this Group* found that the game significantly improved people's ability to detect fake news, how confident they were in their own judgments, and reduced their overall willingness to share misinformation with others (Basol et al., **2022**).

There could be a myriad of explanations for the discrepant results observed. therefore, we categorize explanations into two broad categories: (1) cross-cultural (Indian sample, translated to Hindi) and (2) perspective shift (the player assumed the role of detective).

Firstly, we discuss possible cross-cultural explanations for our observed findings. While inoculation interventions demonstrate a clear potential to be effective (Traberg et al., **2022**), it is not surprising that the process of applying an intervention to understudied, non-WEIRD cultures (Henrich et al., **2010**; Rad et al., **2018**) might require an iterative process. Indeed, previous interventions aiming to reduce belief in and sharing of misinformation in India have faced similar difficulty. WhatsApp's media literacy campaigns and adverts have been criticized for a lack of alignment with local contexts (Medeiros & Singh, **2021**). In-person or online digital literacy interventions have either demonstrated no reduced belief in misinformation (Badrinathan, **2021**) or an effect size limited to a highly educated subset (Guess et al., **2020**). Here, we tested the efficacy of an inoculation intervention, *Join this group,* that was modified for context through partnership with a local non-profit. The intervention aimed to teach participants fundamental techniques commonly used in the presentation of misinformation through an inoculation intervention. We expected that our local adaptation and use of inoculation would improve individual veracity discernment of manipulative news items. Yet, we do not find this in our study.

We hypothesize that the cultural context, local values, and social preferences may have played a role. In particular, the process of successful inoculation in the Indian population may be different. Threat has long been conceptualized a key and necessary component for inoculation to take place (McGuire,

[4]This publication of this data is forthcoming. Once published, it can be made available upon request.

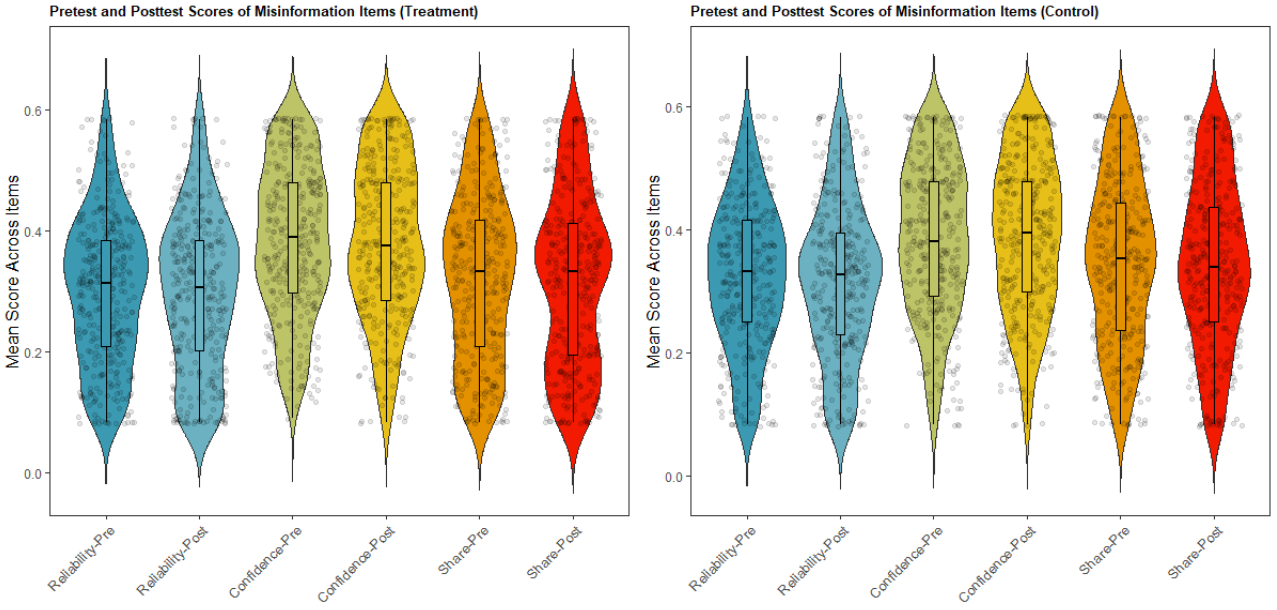


Figure 4 Distribution of pre-test and post-test mean scores in the treatment and control groups, for the reliability, confidence, and sharing scores of misinformation items across all manipulation techniques.

1964) with most recent scholars agreeing that a threshold level of threat is required for inoculation to be conferred (Compton, **2021**) as it serves the function of highlighting one's vulnerability which in turn, motivates the build-up of resistance. While there is no quantitatively defined level of minimum threat discussed in inoculation theory, studies assessing inoculation have traditionally measured threat as an apprehension (Ivanov et al., **2022**; Wood, **2007**) and more recently in a motivational form (Banas & Richards, **2017**). Unfortunately, we did not include measures of apprehensive or motivational threat in our study. Moreover, given the paucity of literature around non-WEIRD samples in psychology in general, it is difficult to make claims about the efficacy of inoculation without an explicit measurement of threat. Future research should consider incorporating this, informed by cultural variation in emotional experience and motivations (Kwan, **2016**; Lim, **2004**; Matsumoto et al., **2008**; Mesquita & Walker, **2003**).

The cross-cultural adaptation also required numerous language and context changes. (Roozenbeek & van der Linden, **2020**). For example, the chosen fictional country of "Santhala" may have carried pre-conceived notions for some given its close resemblance to the Santhal tribe (The Editors of Encyclopaedia Britannica, **2012**). All 12 manipulative WhatsApp prompts were translated from English to Hindi, which may have resulted in a loss of meaning and validity of measurement (see Figure S9 for an example). In addition, based on 2011 national census data, we estimate that our sample is 74% rural (Government of India, **2016**), a figure calculated based on the sample's distribution across states (see Table S39). Shahid et al. (**2022**) find that rural samples had a lower ability to detect misinformation compared to their urban counterparts, suggesting that interventions on rural samples may require additional challenges.

Moreover, rural areas are estimated to have a digital literacy rate of 25% compared to 61% in urban areas (Mothkoor & Mumtaz, **2021**), suggesting that our sample has low digital literacy overall. Classifying a household as digitally literate only requires one person, aged above 5 years, to be able to operate a computer and use the internet. As such, it is likely that our

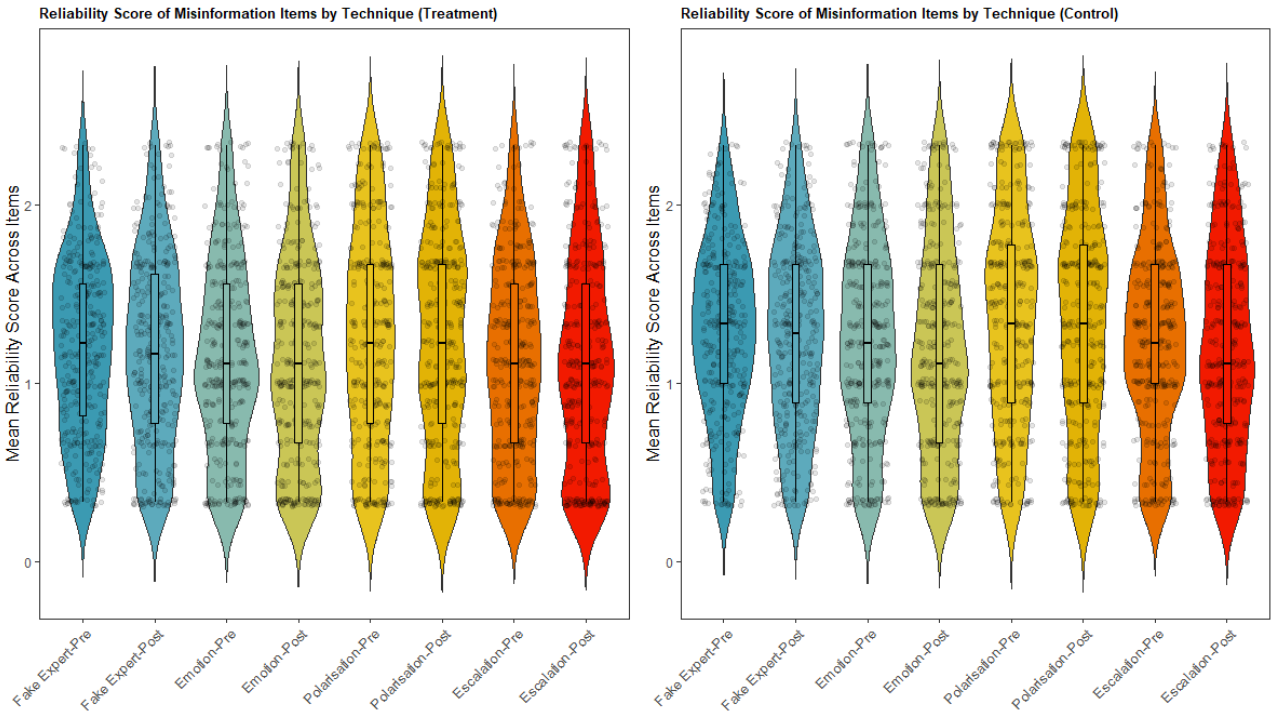


Figure 5 Distribution of mean reliability scores of misinformation items by manipulation technique.

game-based intervention was conducted on participants with minimal experience with operating digital devices. This is compounded by the fact that the majority of our sample was female (55%), who typically have lower digital literacy in this area (Rowntree et al., **2020**). This could have hindered the intervention's efficacy. Furthermore, data quality was poor: only 26% of individuals who played the inoculation game put in the password correctly. Further analysis, however, demonstrated that this did not make a difference to our results (please see Tables S36-38).

Secondly, the game departed from previous game-based inoculation experiments in that it changed the player's perspective from troll to detective. Although this change preserved the critical element of 'active' inoculation that has been effective previously (Pfau et al., **2005**; Roozenbeek & van der Linden, **2019**), it is possible that the role of being not only a detective, but also being undercover, added further layers of complexity that minimized goal salience and clarity for participants, thus reducing its effectiveness. Practitioners may also consider running naturalistic studies in developing countries by conducting interventions broadcasted on WhatsApp through local organizations' subscription lists for increased data availability (Bowles et al., **2020**), or even by artificially constructing a social network in the lab (Pogorelskiy & Shum, **2017**).

Our study may be taken as a lesson in conducting interventions in underexplored populations. In particular, the typical data quality, representativeness, and methodological best practices for running such online experiments in India, and non-WEIRD countries in general, is poorly understood and can impede the experimental process. Campbell-Smith and Bradshaw (**2019**) notes, "having digital connectivity does not mean people are digitally equipped to use online surveys. They have issues in reading and writing, but not in talking." Although we partnered with a local NGO in India, one must also account for gaps in the implementation of scientific experimental designs in the field, particularly by non-academic partners as

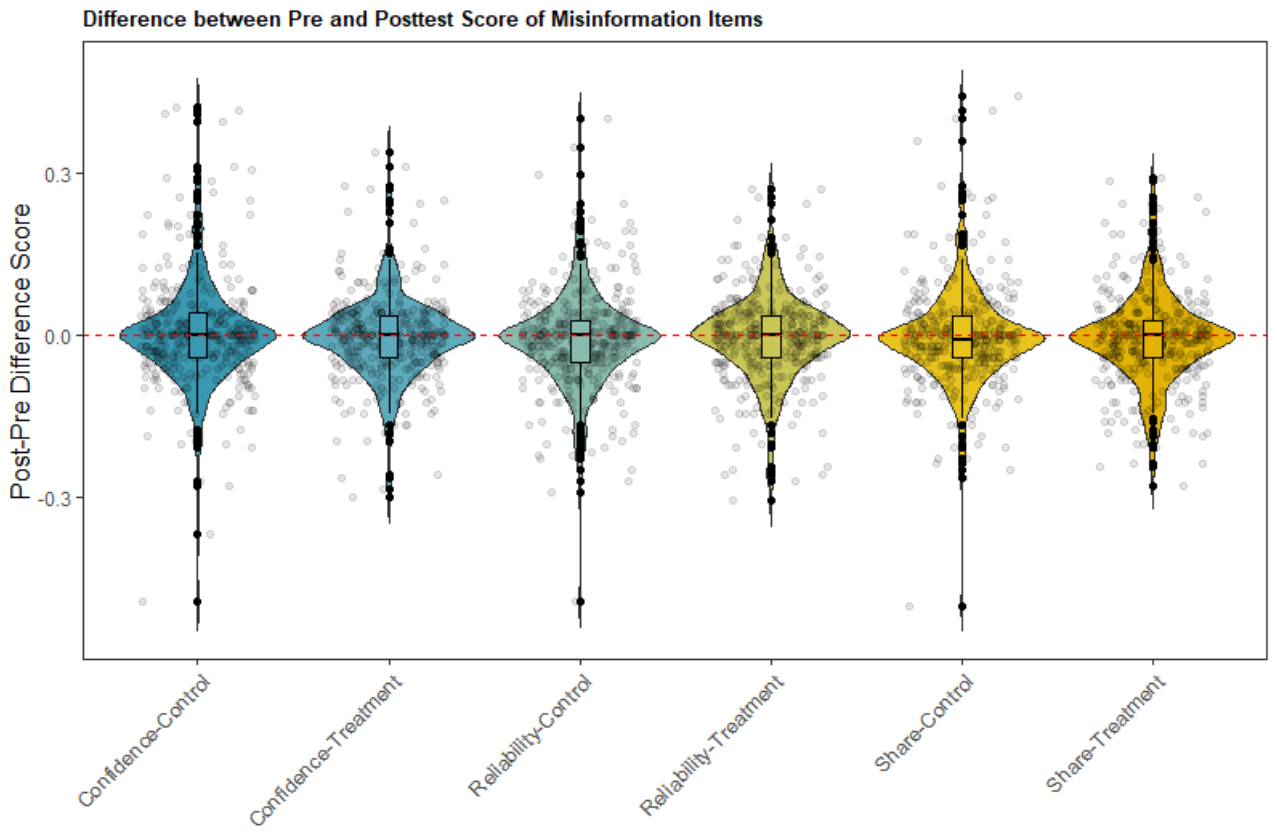


Figure 6 Distribution of post-pre differences between control and treatment groups. Red line drawn at y = 0.

Original Purpose

This paper aims to address the paucity of empirical research investigating misinformation interventions in developing countries. One important difference in the circulation of misinformation in developing countries is its spread through private, encrypted networks such as WhatsApp, which poses different challenges than (the circulation of) misinformation on open networks such as Twitter and Facebook. As such, this paper features a study testing the efficacy of an "inoculation" game in India. We hypothesized that previously reported effects of this inoculation game would be replicated by reducing the reported reliability and sharing intent of misinformation while increasing people's confidence in their own assessments.

it can increase the possibility of unobserved extraneous variables. Additionally, we observed non-random missingness in the data. We find that being assigned to the treatment group increases the odds of an incomplete or missing response, which may have introduced a bias in the results. However, as we found null results no further correction analysis was conducted. Future replications, particularly that find significant results, should pay attention to any differential attrition.

Future studies may also benefit from stronger local relationships (Sircar & Chauchard, **2019**) as well as a greater accountability of the diversity within countries, such as India, that have notable heterogeneity beyond age, gender, and education level (Deshmukh, **2019**). For example, the question on political ideology in this study was more accurately asking people how "free" their

ideology is rather than measuring their political ideology on a left-right scale (measure detailed in the supplement). Although India has been historically classified as clientelist and thus there is no established scale to capture political ideology, some evidence suggests voting behavior among certain groups is not clientelist (Chibber & Verma, **2018**). Future research will need to account for this in the design of surveys. In the context of misinformation, educational interventions have shown differing efficacy depending on political party support (Badrinathan, **2021**) while polarizing content on the basis of religion and caste is often featured in misinformation circulated in India (Al-Zaman, **2021**; Arun, **2019**; Campbell-Smith & Bradshaw, **2019**). For digital interventions, Indian samples may also vary in levels of digital literacy by caste and consumption levels (Mothkoor & Mumtaz, **2021**). Therefore, additional measures, such as whether someone is part of a scheduled group (caste or tribe), religion, income level, and political party affiliation can facilitate a richer understanding of the intervention efficacy in subgroups due to heterogeneity in local factors. To isolate the effect of culture, experiments may also aspire to reach a more digitally literate population within non-WEIRD cultures, given that middle class, urban population in non-WEIRD countries are more likely to resemble the typically studied WEIRD population (Ghai, **2021**).

Conclusion

This study was motivated by scarcity of studies examining non-WEIRD populations in general (Henrich et al., **2010**), and by the lack of research testing the effectiveness of misinformation interventions in democracies such as India (Badrinathan, **2021**), that are being threatened by the prevalence of misinformation. We find null results of a game-based inoculation intervention, *Join this Group,* on ratings of reliability, reported intent to share, and confidence in judgments of misinformation messages. Previous similar game-based inoculation interventions have been demonstrably successful (Basol et al., **2020**; Roozenbeek & van der Linden, **2018**, **2019**, **2020**). We would thus conclude that the results reported here are more likely to reflect an interplay of cultural and experimental design factors. Taken together, we interpret these findings as a call for further adaptation and testing of inoculation interventions on non-WEIRD populations. Modifications may include measuring conceptual mediators such as motivational threat to elucidate and hypothesize potential differences in cross-cultural mechanisms, partnering with local researchers and universities, measuring digital literacy, as well as assessing of behavioral outcomes such as news sharing online.

Acknowledgments

We would like to thank our partners Digital Empowerment Foundation in India for implementing the survey and WhatsApp/Meta for funding.

Funding

This research was funded by WhatsApp through their Research Awards for Social Science and Misinformation program.

References

Adewole, K. S., Anuar, N. B., Kamsin, A., Varathan, K. D., & Razak, S. A. (2017). Malicious accounts: Dark of the social networks. *Journal of Network and Computer Applications*, *79*, 41–67. https://doi.org/10.1016/j.jnca.2016.11.030 (see p. 6).

Al-Zaman, M. S. (2021). A thematic analysis of misinformation in India during the COVID-19 pandemic. *International Information and Library Review*, 1–11. https://doi.org/10.1080/10572317.2021.1908063 (see pp. 1, 13).

Arun, C. (2019). On WhatsApp, rumours, lynchings, and the Indian government. *Economic & Political Weekly*, *54*(6), 30–35. https://papers.ssrn.com/sol3/papers.cfm?abstract%5C_id=3336127 (see pp. 1, 13).

Badrinathan, S. (2021). Educative interventions to combat misinformation: Evidence from a field experiment in india. *American Political Science Review*, *115*(4), 1325–1341. https://doi.org/10.1017/S0003055421000459 (see pp. 1, 2, 9, 13).

Banaji, S., Bhat, R., Agarwal, A., Passanha, N., & Sadhana Pravin, M. (2019). WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India. http://eprints.lse.ac.uk/104316/ (see pp. 1, 2).

Banas, J. A., & Miller, G. (2013). Inducing resistance to conspiracy theory propaganda: Testing inoculation and metainoculation strategies. *Human Communication Research*, *39*(2), 184–207. https://doi.org/10.1111/hcre.12000 (see p. 3).

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311. https://doi.org/10.1080/03637751003758193 (see p. 3).

Banas, J. A., & Richards, A. S. (2017). Apprehension or motivation to defend attitudes? exploring the underlying threat mechanism in inoculation-induced resistance to persuasion. *Communication Monographs*, *84*(2), 164–178. https://doi.org/10.1080/03637751.2017.1307999 (see p. 10).

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 205395172110138. https://doi.org/10.1177/20539517211013868 (see pp. 4, 5, 9).

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1), 2. https://doi.org/10.5334/joc.91 (see pp. 3, 5, 13).

Basol, M., Roozenbeek, J., & van der Linden, S. (2022). Gamified inoculation against misinformation on WhatsApp (see p. 9).

BBC Monitoring. (2021, May 16). *Israel-palestinian conflict: False and misleading claims fact-checked*. https://www.bbc.co.uk/news/57111293 (see pp. 4, 6).

Bhattacharjee, B., Pansari, S., & Dutta, A. (2021). Internet adoption in India. https://images.assettype.com/afaqs/2021-06/b9a3220f-ae2f-43db-a0b4-36a372b243c4/KANTAR%5C_ICUBE%5C_2020%5C_Report%5C_C1.pdf (see p. 2).

Bonetto, E., Troïan, J., Varet, F., Monaco, G., & Girandola, F. (2018). Priming resistance to persuasion decreases adherence to conspiracy theories. *Social Influence*, *13*(3), 125–136. https://doi.org/10.1080/15534510.2018.1471415 (see p. 3).

Bowles, J., Larreguy, H., & Liu, S. (2020). Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in zimbabwe. *PLoS ONE*, *15*(10), 0240005. https://doi.org/10.1371/journal.pone.0240005 (see pp. 2, 11).

Campbell-Smith, U., & Bradshaw, S. (2019). *Global cyber troops country profile: India*. Oxford Internet Institute, University of Oxford. https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/05/India-Profile.pdf (see pp. 2, 11, 13).

Chibber, K. P., & Verma, R. (2018). The myth of cote buying in India. *Ideology and Identity: The Changing Party Systems of India*, 103–130 (see pp. 1, 13).

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, *42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0 (see p. 2).

Compton, J. (2012). Inoculation theory. In *The sage handbook of persuasion: Developments in theory and practice* (pp. 220–236). Sage Publications, Inc. (See p. 3).

Compton, J. (2021). Threat and/in inoculation theory. *International Journal of Communication (Online*, *15*(13), 4294–4307. https://ijoc.org/index.php/ijoc/article/view/17634 (see p. 10).

Compton, J., Linden, S., Cook, J., & Basol, M. (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). https://doi.org/10.1111/spc3.12602 (see p. 3).

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PLOS ONE*, *12*(5), 0175799. https://doi.org/10.1371/journal.pone.0175799 (see p. 3).

Deshmukh, Y. (2019). Methodological issues and problems of conducting surveys in India. a commentary by the Indian issp partner organization. *International Journal of Sociology*, *49*(5-6), 400–411. https://doi.org/10.1080/00207659.2019.1683286 (see p. 12).

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781 (see p. 8).

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://doi.org/10.1038/s44159-021-00006-y (see p. 2).

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by twitter bots? *First Monday*, *25*(6). https://doi.org/10.5210/fm.v25i6.10633 (see p. 1).

Findlay, S. (2019, February 6). *WhatsApp says Indian rules on encryption 'not possible' to meet*. https://www.ft.com/content/9fcfa604-2a0d-11e9-88a4-c32129756dd8 (see p. 4).

Ghai, S. (2021). It's time to reimagine sample diversity and retire the weird dichotomy. *Nature Human Behaviour*, *5*(8), 971–972. https://doi.org/10.1038/s41562-021-01175-9 (see p. 13).

Goga, O., Venkatadri, G., & Gummadi, K. P. (2015–October 30). The doppelgänger bot attack: Exploring identity impersonation in online social networks. *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 141–153. https://doi.org/10.1145/2815675.2815699 (see pp. 4, 6).

Government of India. (2016). *Rural and urban composition of population – census 2011 and 2011*. https://data.gov.in/resource/rural-and-urban-composition-population-census-2001-and-2011 (see p. 10).

Groenendyk, E. (2018). Competing motives in a polarized electorate: Political responsiveness, identity defensiveness, and the rise of partisan antipathy. *Political Psychology*, *39*(S1), 159–171. https://doi.org/10.1111/pops.12481 (see pp. 4, 6).

Gross, K., & Ambrosio, L. D. (2004). Framing emotional response. *Political Psychology*, *25*(1), 1–29. https://www.jstor.org/stable/3792521 (see pp. 4, 6).

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (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 of the United States of America*, *117*, 15536–15545. https://doi.org/10.1073/pnas.1920498117 (see pp. 2, 9).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X (see pp. 2, 9, 13).

Ivanov, B., Rains, S. A., Dillingham, L. L., Parker, K. A., Geegan, S. A., & Barbati, J. L. (2022). The role of threat and counterarguing in therapeutic inoculation. *Southern Communication Journal*, *87*(1), 15–27. https://doi.org/10.1080/1041794X.2021.1983012 (see p. 10).

Iyengar, A., Gupta, P., & Priya, N. (2022). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology*. https://doi.org/10.1002/acp.3995 (see p. 9).

Iyengar, S., & Krupenkin, M. (2018). The strengthening of partisan affect. *Political Psychology*, *39*, 201–218. https://doi.org/10.1111/pops.12487 (see pp. 4, 6).

Jolley, D., & Douglas, K. M. (2017). Prevention is better than cure: Addressing anti-vaccine conspiracy theories. *Journal of Applied Social Psychology*, *47*(8), 459–469. https://doi.org/10.1111/jasp.12453 (see p. 3).

Jung, A. M. (2011). Twittering away the right of publicity: Personality rights and twittering away the right of publicity: Personality rights and celebrity impersonation on social networking websites. *Symposium on Energy Law Article*, *86*(1). https://scholarship.kentlaw.iit.edu/cklawreview/vol86/iss1/16 (see pp. 4, 6).

Kapoor, S., Hughes, P. C., Baldwin, J. R., & Blue, J. (2003). The relationship of individualism–collectivism and self-construals to communication styles in India and the United States. *International Journal of Intercultural Relations*, *27*(6), 683–700. https://doi.org/10.1016/j.ijintrel.2003.08.002 (see p. 2).

Konijn, E. A. (2012). The role of emotion in media use and effects. In *The oxford handbook of media psychology* (pp. 186–211). Oxford University Press. (See pp. 4, 6).

Kwan, L. Y.-Y. (2016). Anger and perception of unfairness and harm: Cultural differences in normative processes that justify sanction assignment. *Asian Journal of Social Psychology*, *19*(1), 6–15. https://doi.org/10.1111/ajsp.12119 (see p. 10).

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963 (see p. 8).

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Era. Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008 (see p. 1).

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018 (see p. 2).

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983 (see p. 3).

Lim, D. H. (2004). Cross cultural differences in online learning motivation. *Educational Media Interna-*

tional, *41*(2), 163–175. https://doi.org/10.1080/09523980410001685784 (see p. 10).

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315 (see p. 3).

Matsumoto, D., Yoo, S. H., & Nakagawa, S. (2008). Culture, emotion regulation, and adjustment. *Journal of Personality and Social Psychology*, *94*(6), 925–937. https://doi.org/10.1037/0022-3514.94.6.925 (see p. 10).

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), 326–332. https://doi.org/10.1037/h0048344 (see p. 3).

McGuire, W. J. (1964). Some contemporary approaches. *Advances in Experimental Social Psychology*, *1*, 191–229. https://doi.org/10.1016/S0065-2601(08)60052-0 (see pp. 3, 9).

McLaughlin, T. (2018, December 12). *How WhatsApp fuels fake news and violence in India*. https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/ (see p. 4).

Medeiros, B., & Singh, P. (2021). Addressing misinformation on WhatsApp in India through intermediary liability policy, platform design modification, and media literacy. *Journal of Information Policy*, *10*, 276–298. https://doi.org/10.5325/JINFOPOLI.10.2020.0276 (see p. 9).

Melki, M., & Pickering, A. (2014). Ideological polarization and the media. *Economics Letters*, *125*(1), 36–39. https://doi.org/10.1016/j.econlet.2014.08.008 (see p. 6).

Mesquita, B., & Walker, R. (2003). Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour Research and Therapy*, *41*(7), 777–793. https://doi.org/10.1016/S0005-7967(02)00189-4 (see p. 10).

Mothkoor, V., & Mumtaz, F. (2021, March 23). *The digital dream: Upskilling India for the future*. https://www.ideasforindia.in/topics/governance/the-digital-dream-upskilling-india-for-the-future.html%5C#:%5Ctextasciitilde%7B%7D:text=Digital%5C%20literacy%5C%20levels%5C%20in%5C%20India%5C&text=Based%5C%20on%5C%20the%5C%20above%5C%20definition,just%5C%2025%5C%25%5C%20in%5C%20rural%5C%20areas. (see pp. 10, 13).

Mumo, M. (2021, August 25). *Protecting burundi's vulnerable media. project syndicate*. https://www.project-syndicate.org/commentary/protecting-press-freedom-in-burundi-by-muthoki-mumo-2021-08 (see p. 1).

Park, A., Bryson, C., Clery, E., Curtice, J., & Philips, M. (2013). *British social attitudes: The 30th report*. NatCen Social Research. https://www.bsa.natcen.ac.uk/latest-report/british-social-attitudes-30/key-findings/introduction.aspx (see p. 5).

Pasquetto, I., Center, S., School, H. K., Jahani, E., Baranovsky, A., & Baum, M. A. (2020). Understanding misinformation on mobile instant messengers (mims) in developing countries. https://shorensteincenter.org/misinformation-on-mims/ (see pp. 1, 2, 4).

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465 (see p. 2).

Pfau, M., Ivanov, B., Houston, B., Haigh, M., Sims, J., Gilchrist, E., Russell, J., Wigley, S., Eckstein, J., & Richert, N. (2005). Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, *72*(4), 414–441. https://doi.org/10.1080/03637750500322578 (see pp. 4, 11).

Pogorelskiy, K., & Shum, M. (2017). News sharing and voting on social networks: An experimental study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2972231 (see p. 11).

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in argentina, nigeria, south africa, and the united kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), 2104235118. https://doi.org/10.1073/pnas.2104235118 (see p. 2).

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115 (see pp. 2, 9).

Rao, M. A., Berry, R., Gonsalves, A., Hastak, Y., Shah, M., & Roeser, R. W. (2013). Globalization and the identity remix among urban adolescents in India. *Journal of Research on Adolescence*, *23*(1), 9–24. https://doi.org/10.1111/jora.12002 (see p. 2).

Rathje, S., Bavel, J. J., & van der Linden, S. (2021). Out-group animosity drives engagement on social media. proceedings of the national academy of

sciences, 118(26, 2024292118. https://doi.org/10.1073/pnas.2024292118 (see p. 6).

Reis, J. C. S., Melo, P., Garimella, K., & Benevenuto, F. (2020). Can WhatsApp benefit from debunked fact-checked stories to reduce misinformation? *Harvard Kennedy School Misinformation Review*, *1*(5). https://doi.org/10.37016/mr-2020-035 (see pp. 1, 2).

Reznik, M. (2013). Identity theft on social networking sites: Developing issues of internet impersonation. *Touro Law Review*, *29*(2), 455–483. https://digitalcommons.tourolaw.edu/lawreviewAvailable at:https://digitalcommons.tourolaw.edu/lawreview/vol29/iss2/12https://digitalcommons.tourolaw.edu/lawreview/vol29/iss2/12 (see pp. 4, 6).

Robb, A. (2021). Anatomy of a fake news scandal. rolling stone. https://www.rollingstone.com/feature/ anatomy-of-a-fake-news-scandal-125877/ (see pp. 4, 6).

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online misinformation. *Solomon Revisited. Educational and Psychological Measurement*, *81*(2), 340–362. https:/ / doi.org/ 10.1177/ 0013164420940378 (see pp. 3, 4).

Roozenbeek, J., & van der Linden, S. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491 (see pp. 3, 13).

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 65. https://doi.org/10.1057/s41599-019-0279-9 (see pp. 2, 3, 8, 11, 13).

Roozenbeek, J., & van der Linden, S. (2020). Breaking harmony square: A game that "inoculates" against political misinformation. *The Harvard Kennedy School (HKS) Misinformation Review*, *1*(8). https://doi.org/ 10.37016/ mr-2020-4 (see pp. 4, 5, 9, 10, 13).

Rowntree, O., Shahnan, M., Bahia, K., Butler, C., Lindsey, D., & Sibthorpe, C. (2020). The mobile gender gap report 2020. https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2020/05/GSMA-The-Mobile-Gender-Gap-Report-2020.pdf (see p. 11).

Sagarin, B. J., Cialdini, R. B., Rice, W. E., & Serna, S. B. (2002). Dispelling the illusion of invulnerability: The motivations and mechanisms of resistance to persuasion. *Journal of Personality and Social Psychology*, *83*(3), 526–541. https://doi.org/10.1037/0022-3514.83.3.526 (see p. 3).

Shahid, F., Mare, S., & Vashistha, A. (2022). Examining source effects perceptions of fake news in India. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSW1, 89), 1–29. https://doi.org/10.1145/3512936 (see p. 10).

Sircar, N., & Chauchard, S. (2019). Dilemmas and challenges of citizen information campaigns: Lessons from a failed experiment in India. *Information, Accountability, and Cumulative Learning*, 287–312. https://doi.org/10.1017/9781108381390.011 (see p. 12).

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, *26*(6), 301–319. https://doi.org/10.1093/jcmc/zmab010 (see p. 1).

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1948–1961. https://doi.org/10.1037/xlm0000422 (see p. 2).

The Editors of Encyclopaedia Britannica. (2012). Santhal. *Encyclopaedia Britannica*. https://www.britannica.com/topic/Santhal (see p. 10).

Traberg, C., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, *700* (see p. 9).

Vaishnav, M., Jaffrelot, C., Mehta, G., Rej, A., Shrinivasan, R., Sagar, R., & Verma, R. (2019). *The bjp in power: Indian democracy and religious nationalism*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2019/04/04/bjp-in-power-indian-democracy-and-religious-nationalism-pub-78677 (see p. 2).

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6 (see pp. 2, 3).

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), 1600008. https:/ / doi.org/10.1002/gch2.201600008 (see p. 3).

van der Linden, S., Maibach, E., Cook, J., Leiserowitz, A., & Lewandowsky, S. (2017). Inoculating against misinformation (J. Sills, Ed.). *Science*, *358*(6367), 1141–1142. https://doi.org/10.1126/science.aar4533 (see p. 1).

Vasudeva, F., & Barkdull, N. (2020). WhatsApp in India? a case study of social media related lynchings.

Social Identities, *26*(5), 574–589. https://doi-org.ezp.lib.cam.ac.uk/10.1080/13504630.2020.1782730 (see p. 1).

Verma, J., & Triandis, H. C. (2020). The measurement of collectivism in India. *Merging Past, Present, and Future in Cross-Cultural Psychology: Selected papers from the Fourteenth International Congress of the International Association for Cross-Cultural Psychology*, 256–265 (see p. 2).

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, *37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894 (see p. 2).

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564 (see p. 2).

Wood, M. L. M. (2007). Rethinking the inoculation analogy: Effects on subjects with differing preexisting attitudes. *Human Communication Research*, *33*(3), 357–378. https://doi.org/10.1111/j.1468-2958.2007.00303.x (see p. 10).

Zollo, F., Novak, P. K., Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., Quattrociocchi, W., & Preis, T. (2015). Emotional dynamics in the age of misinformation. *PLoS ONE*, *10*(9), 0138740. https://doi.org/10.1371/journal.pone.0138740 (see p. 4).

Supplemental Materials

Missing Data

A total of *n* = 1283 consenting individuals began the survey of which *n* = 757 were complete and valid responses used in the analysis. As sample demographics were only collected after the post-test measures, it is not possible to understand the differences in individual characteristics across missing and complete responses. However, after filtering out for the those answered at least one question in the pre-test (*n* = 1038), Little's MCAR test (run in R using the *misty* package) for all three dependent variables (reliability, confidence and sharing) suggested that the data were not missing completely at random, χ^2 (5) = 70.59, $p < 0.001$. Thus, we ran a standard logistic regression (using the *glm* function from the *stats* package in R) to investigate patterns of missing data as a function of pre-test responses. This was done by creating a dummy variable where 1 = missing observation and 0 = complete responses. For the manipulative items, higher pre-test confidence scores slightly reduced the odds of missingness (*OR* = 0.030, [95%*CI*; 0.002,0.431]) and being assigned the treatment group increased the odds of missingness (*OR* = 2.171, [95%*CI*; 1.589, 2.967]). This implies that a higher baseline confidence in assessing the reliability of manipulative items decreases the likelihood of missingness while being assigned to the treatment group increases the likelihood of missingness. All other pre-test measures did not affect the odds of dropout. We were not able to assess whether the missing data was due to demographic factors as these were collected at the end of the study.

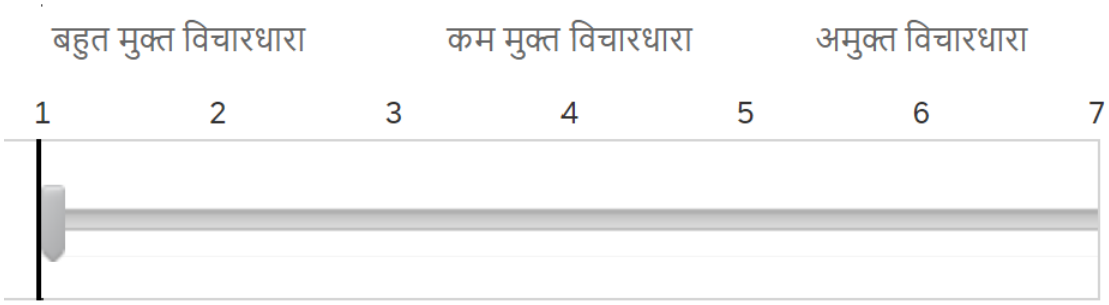
Table S1 Logistic Regression Predicting Missingness (where Missing data = 1, Complete data = 0)

	Odds Ratio	Confidence Intervals
(Intercept)	0.408	*CI* [0.142, 1.172]
Reliability Pre-test (Fake Items)	0.836	*CI* [0.082, 8.490]
Confidence Pre-test (Fake Items)	0.030 **	*CI* [0.002, 0.431]
Sharing Pre-test (Fake Items)	4.965	*CI* [0.509, 48.440]
WhatsApp Usage	1.032	*CI* [0.848, 1.256]
Reliability Pre-test (Real Items)	1.100	*CI* [0.849, 1.425]
Confidence Pre-test (Real Items)	0.811	*CI* [0.626, 1.053]
Sharing Pre-test (Real Items)	1.034	*CI* [0.806, 1.327]
Reliability Pre-test (Real Fake Items)	1.067	*CI* [0.800, 1.424]
Confidence Pre-test (Real Fake Items)	0.874	*CI* [0.660, 1.157]
Sharing Pre-test (Real Fake Items)	1.307	*CI* [0.942, 1.813]
Condition (Treatment)	2.171 ***	*CI* [1.589, 2.967]
N		899
AIC		1057.126
BIC		1114.742
Pseudo R2		0.083

*** p < 0.001; ** p < 0.01; * p < 0.05.

Political Ideology Measurements.

Although we employed a measure from the British Social Attitudes survey, we employed a measure to assess the self-reported identification along the left to right spectrum:



On the slider below, please indicate your political ideology.
[Far left of slider; closer to 1] Very free ideology
[Middle of slider; closer to 4] Less free ideology
[Far right of slider; closer to 7] Not free ideology

Table S2 Sample Composition

Variable	*n*	*Percentage*	*Cumulative Percentage*
Gender			
Male	293	40%	40%
Female	397	55%	95%
Other	35	5%	100%
Age			
18-24	356	49%	49%
25-34	286	39%	89%
35-44	64	9%	97%
45-54	16	2%	100%
55 and over	3	0%	100%
Political Leaning			
1 Very left-wing	139	19%	19%
2	385	53%	72%
3	165	23%	95%
4	34	5%	100%
5 Very right-wing	2	0%	100%
Education			
Class 12	159	22%	22%
Elementary	16	2%	24%
Graduate	306	42%	66%
Post Grad	172	24%	90%
Up to Tenth	72	10%	100%
State			
Bihar	19	3%	3%
Chhattisgarh	42	6%	8%
Delhi	3	0%	9%
Haryana	5	1%	10%
Jharkhand	26	4%	13%
Madhya Pradesh	471	65%	78%
Rajasthan	120	17%	95%
Unknown	6	1%	95%
Uttar Pradesh	33	5%	100%
Frequency of Checking the News			
1 Never	5	1%	1%
2 Occasionally	90	12%	13%
3 Somewhat	166	23%	36%
4 Often	295	41%	77%
5 All the time	169	23%	100%

Table S2 Table S2 continued

Use of social media			
1 Never	28	4%	4%
2 Occasionally	129	18%	22%
3 Somewhat	167	23%	45%
4 Often	212	29%	74%
5 All the time	189	26%	100%
Use of WhatsApp			
1 Never	4	1%	1%
2 Occasionally	22	3%	4%
3 Once a week	26	4%	7%
4 Daily	90	12%	20%
5 More than once a day	520	72%	91%
NA	63	9%	100%
Interest in Politics			
1 Not interested at all	50	7%	7%
2	84	12%	18%
3 Slightly interested	289	40%	58%
4	189	26%	84%
5 Very interested	113	16%	100%

Table S3 ANCOVA on Post-Treatment scores of reliability assessments (of manipulative items)

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.25	1	0.25	36.90	$p < 0.001$		
F_Rel_Pre	6.76	1	6.76	1000.84	$p < 0.001$	.58	[.55, .61]
Condition	0.00	1	0.00	0.00	.969	.00	[.00, 1.00]
Error	4.88	722	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S4 ANCOVA on Post-Treatment scores of confidence measure (of manipulative items)

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.59	1	0.59	83.25	$p < 0.001$		
F_Conf_Pre	6.45	1	6.45	908.30	$p < 0.001$	.56	[.52, .59]
Condition	0.01	1	0.01	1.79	.181	.00	[.00, .01]
Error	5.13	722	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S5 ANCOVA on Post-Treatment scores of sharing measure (of manipulative items)

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.35	1	0.35	48.64	$p < 0.001$		
F_Share_Pre	8.42	1	8.42	1155.91	$p < 0.001$	.62	[.58, .64]
Condition	0.01	1	0.01	1.46	.227	.00	[.00, .01]
Error	5.26	722	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S6 *Bayesian paired sample t-test on dependent variables*

Variable			Statistic	Error %
Reliability of Fake Messages				
Reliability-Post	Reliability-Pre	$BF_{10,\ prior\ =\ 0.707}$	0.249	2.604E^{-08}
Confidence in judgement of Fake Messages				
Confidence-Post	Confidence-Pre	$BF_{10,\ prior\ =\ 0.707}$	0.043	1.612E^{-07}
Intent to share Fake Messages				
Share-Post	Share-Pre	$BF_{10,\ prior\ =\ 0.707}$	0.073	9.425E^{-08}

Table S7 Reliability measure - Fixed-Effects ANCOVA on post-test fake expert manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	6.49	1	6.49	48.51	$p < 0.001$		
FE_Rel_Pre	102.89	1	102.89	768.80	$p < 0.001$	.52	[.48, .55]
Condition	0.03	1	0.03	0.21	.648	.00	[.00, .01]
Error	96.62	722	0.13				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S8 Reliability Measure - ANCOVA on post-test score of emotional manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	14.19	1	14.19	69.20	$p < 0.001$		
EM_Rel_Pre	89.90	1	89.90	438.44	$p < 0.001$	.38	[.33, .42]
Condition	0.04	1	0.04	0.21	.649	.00	[.00, .01]
Error	148.04	722	0.21				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S9 Reliability Measure - ANCOVA on post-test score of polarisation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	16.52	1	16.52	81.00	$p < 0.001$		
PL_Rel_Pre	119.46	1	119.46	585.71	$p < 0.001$	.45	[.41, .49]
Condition	0.07	1	0.07	0.35	.553	.00	[.00, .01]
Error	147.26	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S10 Reliability Measure - ANCOVA on post-test score of escalation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	13.38	1	13.38	68.51	$p < 0.001$		
ES_Rel_Pre	96.59	1	96.59	494.40	$p < 0.001$	.41	[.36, .45]
Condition	0.01	1	0.01	0.03	.852	.00	[.00, .00]
Error	141.06	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S11 *Reliability measure – ANCOVA on post-test score of authentic fake news items*

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	63.45	1	63.45	116.01	.000		
RF_Rel_Pre	198.77	1	198.77	363.43	.000	.34	[.29, .38]
Condition	0.54	1	0.54	0.99	.319	.00	[.00, .01]
Error	389.40	712	0.55				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S12 Reliability measure - ANCOVA on post-test score of real (non-manipulative) items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	41.14	1	41.14	81.48	.000		
R_Rel_Pre	265.29	1	265.29	525.46	.000	.42	[.38, .46]
Condition	0.05	1	0.05	0.09	.763	.00	[.00, .00]
Error	364.52	722	0.50				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S13 Confidence measure - ANCOVA on post-test score of fake expert manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	16.56	1	16.56	123.94	$p < 0.001$		
FE_Conf_Pre	96.67	1	96.67	723.45	$p < 0.001$	.50	[.46, .54]
Condition	0.21	1	0.21	1.56	.211	.00	[.00, .01]
Error	96.47	722	0.13				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S14 Confidence measure – ANCOVA on post-test score of emotional manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	19.49	1	19.49	103.28	$p < 0.001$		
EM_Conf_Pre	90.92	1	90.92	481.79	$p < 0.001$	.40	[.36, .44]
Condition	0.20	1	0.20	1.05	.306	.00	[.00, .01]
Error	136.25	722	0.19				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S15 Confidence measure – ANCOVA on post-test score of polarisation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	22.17	1	22.17	119.57	$p < 0.001$		
PL_Conf_Pre	92.32	1	92.32	497.93	$p < 0.001$	.41	[.37, .45]
Condition	0.22	1	0.22	1.18	.278	.00	[.00, .01]
Error	133.87	722	0.19				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S16 Confidence measure – ANCOVA on post-test score of escalation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	24.97	1	24.97	135.70	$p < 0.001$		
ES_Conf_Pre	90.68	1	90.68	492.84	$p < 0.001$	.41	[.36, .45]
Condition	0.21	1	0.21	1.17	.280	.00	[.00, .01]
Error	132.84	722	0.18				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S17 Confidence measure – ANCOVA on post-test score of authentic fake news items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\eta^2$	$_{partial}\eta^2$ 90%*CI* [LL, UL]
(Intercept)	72.95	1	72.95	143.40	.000		
RF_Conf_Pre	200.25	1	200.25	393.60	.000	.36	[.31, .40]
Condition	0.86	1	0.86	1.68	.195	.00	[.00, .01]
Error	361.73	711	0.51				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S18 Confidence measure – ANCOVA on post-test score of real (non-manipulative) items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	64.69	1	64.69	110.28	.000		
R_Conf_Pre	325.73	1	325.73	555.32	.000	.43	[.39, .47]
Condition	0.65	1	0.65	1.10	.295	.00	[.00, .01]
Error	423.49	722	0.59				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S19 Sharing Measure – ANCOVA on post-test score of fake expert manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	10.08	1	10.08	69.52	$p < 0.001$		
FE_Share_Pre	125.70	1	125.70	867.08	$p < 0.001$	.55	[.51, .58]
Condition	0.28	1	0.28	1.94	.164	.00	[.00, .01]
Error	104.67	722	0.14				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S20 Sharing Measure – ANCOVA on post-test score of emotional manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	13.68	1	13.68	67.67	$p < 0.001$		
EM_Share_Pre	133.09	1	133.09	658.41	$p < 0.001$	.48	[.44, .51]
Condition	0.06	1	0.06	0.29	.590	.00	[.00, .01]
Error	145.95	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S21 Sharing Measure – ANCOVA on post-test score of polarisation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	17.68	1	17.68	82.23	$p < 0.001$		
PL_Share_Pre	130.23	1	130.23	605.59	$p < 0.001$	.46	[.41, .49]
Condition	0.59	1	0.59	2.75	.098	.00	[.00, .01]
Error	155.26	722	0.22				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S22 Sharing Measure – ANCOVA on post-test score of escalation manipulation items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	17.43	1	17.43	88.22	$p < 0.001$		
ES_Share_Pre	130.11	1	130.11	658.64	$p < 0.001$	.48	[.44, .51]
Condition	0.55	1	0.55	2.77	.097	.00	[.00, .01]
Error	142.63	722	0.20				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S23 Sharing measure – ANCOVA on post-test score of authentic fake news items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	49.16	1	49.16	83.26	.000		
RF_Share_Pre	291.58	1	291.58	493.82	.000	.41	[.37, .45]
Condition	0.07	1	0.07	0.12	.732	.00	[.00, .00]
Error	414.50	702	0.59				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S24 Sharing Measure – ANCOVA on post-test score of real (non-manipulative) items

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	44.18	1	44.18	78.11	.000		
R_Share_Pre	373.52	1	373.52	660.31	.000	.48	[.44, .51]
Condition	0.79	1	0.79	1.39	.239	.00	[.00, .01]
Error	408.41	722	0.57				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S25 Pre-Post Mean Differences

Variable	*Condition*	*N*	*Mean.Difference*	*SD*
Reliability (manipulative items)	Treatment	360	-0.00	0.08
Confidence (manipulative items)	Treatment	360	-0.00	0.08
Sharing (manipulative items)	Treatment	360	-0.00	0.08
Reliability (real items)	Treatment	360	0.07	0.73
Confidence (real items)	Treatment	360	-0.02	0.83
Sharing (real items)	Treatment	360	0.05	0.80
Reliability (authentic fake items)	Treatment	355	0.04	0.78
Confidence (authentic fake items)	Treatment	357	-0.01	0.78
Sharing (authentic fake items)	Treatment	352	0.01	0.84
Reliability (manipulative items)	Control	365	-0.01	0.09
Confidence (manipulative items)	Control	365	0.00	0.10
Sharing (manipulative items)	Control	365	-0.00	0.09
Reliability (real items)	Control	365	0.04	0.80
Confidence (real items)	Control	365	0.04	0.84
Sharing (real items)	Control	365	0.07	0.81
Reliability (authentic fake items)	Control	360	0.03	0.85
Confidence (authentic fake items)	Control	357	0.04	0.81
Sharing (authentic fake items)	Control	353	-0.03	0.84

Table S26 Reliability Measure - item-level ANOVA table (pre-post difference scores)

Variable	*F.value*	*df1*	*df2*	*p*
Diff_Fake_Rel_1-FakeExp	0.271	1	723	0.603
Diff_Fake_Rel_2-FakeExp	0.108	1	723	0.743
Diff_Fake_Rel_3-FakeExp	0.303	1	723	0.582
Diff_Fake_Rel_4-Emotion	0.044	1	723	0.834
Diff_Fake_Rel_5-Emotion	3.286	1	723	0.070
Diff_Fake_Rel_6-Polarise	0.371	1	723	0.543
Diff_Fake_Rel_7-Emotion	1.407	1	723	0.236
Diff_Fake_Rel_8-Polarise	1.744	1	723	0.187
Diff_Fake_Rel_9-Polarise	0.010	1	723	0.919
Diff_Fake_Rel_10-Escalate	0.322	1	723	0.571
Diff_Fake_Rel_11-Escalate	1.317	1	723	0.252
Diff_Fake_Rel_12-Escalate	0.008	1	723	0.930
Diff_Real_Fake_Conf_13 (authentic fake item)	0.186	1	713	0.666
Diff_Real_Fake_Conf_14 (authentic fake item)	0.283	1	723	0.595
Diff_Real_Rel_15	0.191	1	723	0.662
Diff_Real_Rel_16	1.345	1	723	0.247

Table S27 Confidence Measure - item-level ANOVA table (pre-post difference scores)

Variable	*F.value*	*df1*	*df2*	*p*
Diff_Fake_Conf_1-FakeExp	2.323	1	723	0.128
Diff_Fake_Conf_2-FakeExp	0.001	1	723	0.974
Diff_Fake_Conf_3-FakeExp	0.000	1	723	0.990
Diff_Fake_Conf_4-Emotion	0.214	1	723	0.644
Diff_Fake_Conf_5-Emotion	0.576	1	723	0.448
Diff_Fake_Conf_6-Polarise	2.496	1	723	0.115
Diff_Fake_Conf_7-Emotion	0.327	1	723	0.567
Diff_Fake_Conf_8-Polarise	1.697	1	723	0.193
Diff_Fake_Conf_9-Polarise	3.400	1	723	0.066
Diff_Fake_Conf_10-Escalate	0.035	1	723	0.851
Diff_Fake_Conf_11-Escalate	0.929	1	723	0.336
Diff_Fake_Conf_12-Escalate	0.807	1	723	0.369
Diff_Real_Fake_Conf_13 (authentic fake item)	2.458	1	712	0.117
Diff_Real_Fake_Conf_14 (authentic fake item)	0.337	1	723	0.562
Diff_Real_Conf_15	2.044	1	723	0.153
Diff_Real_Conf_16	0.002	1	723	0.965

Table S28 Sharing Measure - item-level ANOVA table (pre-post difference scores)

Variable	*F.value*	*df1*	*df2*	*p*
Diff_Fake_Share_1-FakeExp	0.592	1	723	0.442
Diff_Fake_Share_2-FakeExp	0.385	1	723	0.535
Diff_Fake_Share_3-FakeExp	0.004	1	723	0.952
Diff_Fake_Share_4-Emotion	0.012	1	723	0.911
Diff_Fake_Share_5-Emotion	1.179	1	723	0.278
Diff_Fake_Share_6-Polarise	0.233	1	723	0.629
Diff_Fake_Share_7-Emotion	0.010	1	723	0.921
Diff_Fake_Share_8-Polarise	1.426	1	723	0.233
Diff_Fake_Share_9-Polarise	0.672	1	723	0.413
Diff_Fake_Share_10-Escalate	2.935	1	723	0.087
Diff_Fake_Share_11-Escalate	0.146	1	723	0.703
Diff_Fake_Share_12-Escalate	0.144	1	723	0.705
Diff_Real_Fake_Conf_13 (authentic fake item)	0.664	1	703	0.415
Diff_Real_Fake_Conf_14 (authentic fake item)	0.099	1	723	0.753
Diff_Real_Share_15	0.006	1	723	0.936
Diff_Real_Share_16	0.203	1	723	0.653

Table S29 Reliability Measure – Item-level statistics

	Treatment				**Control**			
Item	*Mpre*	*SDpre*	*Mpost*	*SDpost*	*Mpre*	*SDpre*	*Mpost*	*SDpost*
Fake_Rel_1-FakeExp	4.51	2.27	4.28	2.30	4.50	2.27	4.36	2.26
Fake_Rel_10-Escalate	3.69	2.17	3.50	2.18	4.11	2.24	3.83	2.27
Fake_Rel_11-Escalate	2.95	1.98	3.09	2.07	3.24	2.13	3.20	2.08
Fake_Rel_12-Escalate	3.59	2.15	3.55	2.13	3.84	2.15	3.82	2.10
Fake_Rel_2-FakeExp	3.12	2.08	3.19	2.13	3.40	2.16	3.42	2.14
Fake_Rel_3-FakeExp	3.92	2.26	3.77	2.21	4.18	2.29	4.12	2.24
Fake_Rel_4-Emotion	3.44	1.97	3.36	2.05	3.63	2.09	3.59	2.10
Fake_Rel_5-Emotion	3.68	2.13	3.61	2.12	4.15	2.14	3.78	2.09
Fake_Rel_6-Polarise	4.03	2.26	3.87	2.25	4.21	2.20	4.16	2.22
Fake_Rel_7-Emotion	3.24	2.07	3.31	2.16	3.50	2.05	3.38	2.15
Fake_Rel_8-Polarise	3.53	2.06	3.61	2.18	3.95	2.06	3.82	2.14
Fake_Rel_9-Polarise	3.79	2.21	3.82	2.17	4.07	2.18	4.12	2.17
Real_Fake Rel_13	3.26	2.06	3.36	2.06	3.45	2.03	3.65	2.06
Real_Fake Rel_14	3.32	2.18	3.34	2.28	3.76	2.29	3.69	2.28
Real_Rel_15	2.49	1.93	2.59	1.98	2.77	2.10	2.93	2.13
Real_Rel_16	2.65	2.00	2.82	2.11	2.89	2.06	2.89	2.04

Table S30 Confidence Measure – Item-level statistics

	Treatment				Control			
Item	*Mpre*	*SDpre*	*Mpost*	*SDpost*	*Mpre*	*SDpre*	*Mpost*	*SDpost*
Fake_Conf_1-FakeExp	4.80	2.01	4.84	1.94	4.83	2.12	5.09	1.93
Fake_Conf_10-Escalate	4.57	2.02	4.56	1.96	4.68	2.00	4.69	1.98
Fake_Conf_11-Escalate	4.51	2.04	4.50	2.07	4.42	2.19	4.56	2.05
Fake_Conf_12-Escalate	4.60	2.01	4.49	1.93	4.57	2.00	4.60	1.96
Fake_Conf_2-FakeExp	4.32	2.03	4.36	2.04	4.47	2.12	4.50	2.10
Fake_Conf_3-FakeExp	4.66	2.04	4.72	1.94	4.72	2.07	4.79	2.00
Fake_Conf_4-Emotion	4.32	1.99	4.24	1.97	4.42	2.01	4.41	2.03
Fake_Conf_5-Emotion	4.66	1.96	4.54	1.95	4.67	2.04	4.66	1.92
Fake_Conf_6-Polarise	4.85	1.95	4.63	1.98	4.78	2.01	4.81	2.05
Fake_Conf_7-Emotion	4.44	1.98	4.36	2.02	4.39	1.99	4.41	1.97
Fake_Conf_8-Polarise	4.54	1.94	4.64	2.01	4.72	1.95	4.61	1.97
Fake_Conf_9-Polarise	4.71	1.97	4.66	1.99	4.55	2.01	4.79	1.89
Real Fake_Conf_13	4.54	2.02	4.46	2.00	4.55	1.98	4.69	1.87
Real Fake_Conf_14	4.45	2.11	4.50	2.13	4.70	2.14	4.66	2.05
Real_Conf_15	4.41	2.27	4.25	2.25	4.29	2.24	4.35	2.21
Real_Conf_16	4.21	2.26	4.29	2.18	4.40	2.19	4.48	2.20

Table S31 Sharing Measure – Item-level statistics

	Treatment				Control			
Item	*Mpre*	*SDpre*	*Mpost*	*SDpost*	*Mpre*	*SDpre*	*Mpost*	*SDpost*
Fake_Share_1-FakeExp	4.55	2.28	4.39	2.25	4.73	2.26	4.70	2.21
Fake_Share_10-Escalate	4.03	2.24	3.79	2.24	4.32	2.27	4.35	2.21
Fake_Share_11-Escalate	3.37	2.19	3.51	2.21	3.65	2.31	3.73	2.27
Fake_Share_12-Escalate	3.93	2.25	3.83	2.26	4.12	2.17	4.08	2.24
Fake_Share_2-FakeExp	3.57	2.23	3.57	2.18	3.91	2.28	3.82	2.23
Fake_Share_3-FakeExp	3.92	2.26	4.03	2.31	4.30	2.28	4.42	2.25
Fake_Share_4-Emotion	3.58	2.16	3.51	2.19	3.88	2.29	3.83	2.20
Fake_Share_5-Emotion	3.88	2.19	3.80	2.19	4.33	2.21	4.08	2.18
Fake_Share_6-Polarise	4.03	2.31	4.00	2.32	4.55	2.24	4.45	2.23
Fake_Share_7-Emotion	3.38	2.17	3.45	2.23	3.68	2.18	3.77	2.23
Fake_Share_8-Polarise	3.93	2.24	3.78	2.26	4.11	2.18	4.16	2.27
Fake_Share_9-Polarise	4.14	2.22	3.92	2.29	4.29	2.20	4.21	2.14
Real Fake_Share_13	3.62	2.26	3.75	2.23	4.03	2.31	4.04	2.22
Real Fake Share_14	3.84	2.27	3.79	2.29	4.19	2.37	4.09	2.33
Real_Share_15	2.94	2.12	3.05	2.20	3.48	2.28	3.58	2.33
Real_Share_16	3.20	2.30	3.31	2.30	3.40	2.29	3.58	2.34

Table S32 Reliability, Confidence and Sharing Measure of all manipulative items - Two-sided Independent Samples t-test of equivalence (TOSTs)

Var	*b.0.*	*t.0.*	*df.0.*	*p.0.*	*b.1.*	*t.1.*	*df.1.*	*p.1.*	*b.2.*	*t.2.*	*df.2.*	*p.2.*
F_Rel_Post	t-test	1.92	721.68	0.06	TOST Upper	-1.45	721.68	0.07	TOST Lower	5.29	721.68	$p < 0.001$
F_Conf_Post	t-test	1.03	721.43	0.31	TOST Upper	-2.34	721.43	0.01	TOST Lower	4.39	721.43	$p < 0.001$
F_Share_Post	t-test	2.72	719.73	0.01	TOST Upper	-0.64	719.73	0.26	TOST Lower	6.09	719.73	$p < 0.001$

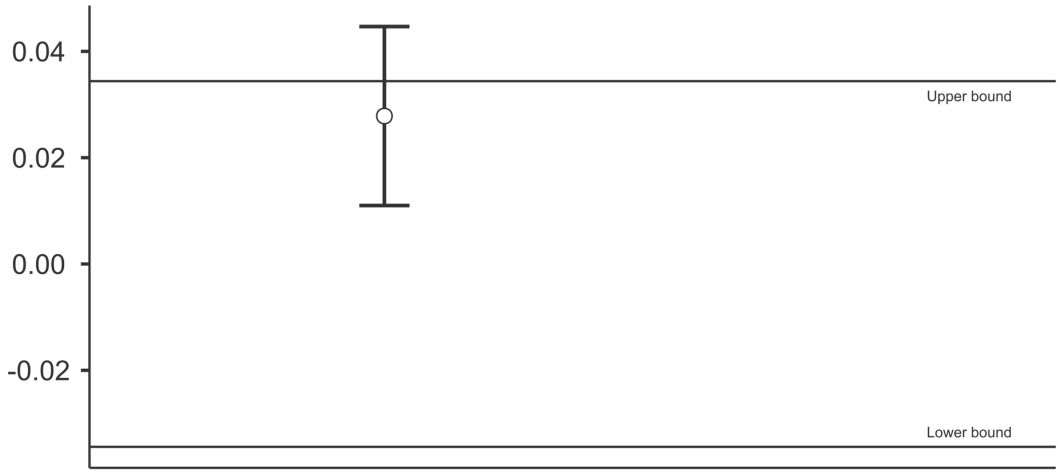


Table S33 Linear regression with difference in pre-post reliability rating of manipulative messaging as the dependent variable

Predictors	*Estimates*	CI	*p*
(Intercept)	−0.02	−0.08 – −0.03	0.436
Condition [Treatment]	0.00	−0.01 – −0.02	0.633
Gender [2]	−0.01	−0.02 – −0.00	0.186
Gender [3]	−0.03	−0.06 – −0.00	0.069
Grad [1]	−0.00	−0.01 – −0.01	0.955
Age25-34	−0.00	−0.02 – −0.01	0.580
Age35-44	0.01	−0.02 – −0.03	0.542
Age45-54	−0.03	−0.08 – −0.01	0.130
Age [55 and over]	0.00	−0.09 – −0.10	0.964
Pol_interest_1	−0.00	−0.01 – −0.01	0.973
LR_Score	−0.00	−0.01 – −0.01	0.906
FromMP [1]	0.01	−0.01 – −0.02	0.517
Lib_Auth	0.01	−0.00 – −0.03	0.054
WAUse_1	−0.01	−0.01 – −0.00	0.198
News.checking_1	0.01	−0.00 – −0.01	0.087
Social.checking_1	−0.00	−0.01 – −0.00	0.527
Observations		662	
R^2 / R^2 adjusted		0.026 / 0.003	

Table S34 Linear regression with difference in pre-post confidence rating of manipulative messaging as the dependent variable

Predictors	*Estimates*	CI %	*p*
(Intercept)	−0.01	−0.07 – −0.05 %	0.783
Condition [Treatment]	−0.01	−0.02 – −0.01 %	0.335
Gender [2]	−0.00	−0.02 – −0.01 %	0.570
Gender [3]	−0.01	−0.04 – −0.02 %	0.622
Grad [1]	−0.01	−0.02 – −0.01 %	0.218
Age25-34	−0.01	−0.02 – −0.01 %	0.492
Age35-44	−0.01	−0.03 – −0.02 %	0.603
Age45-54	−0.04	−0.08 – −0.01 %	0.103
Age [55 and over]	0.00	−0.10 – −0.10 %	0.944
Pol_interest_1	−0.00	−0.01 – −0.00 %	0.421
LR_Score	−0.01	−0.02 – −0.00 %	0.310
FromMP [1]	−0.01	−0.02 – −0.01 %	0.523
Lib_Auth	0.01	−0.01 – −0.02 %	0.253
WAUse_1	0.00	−0.01 – −0.01 %	0.611
News.checking_1	0.01	0.00 – −0.02 %	**0.032**
Social.checking_1	−0.00	−0.01 – −0.00 %	0.203
Observations		662	
R^2 / R^2 adjusted		0.019 / -0.003	

Table S35 Linear regression with difference in pre-post sharing rating of manipulative messaging as the dependent variable

Predictors	*Estimates*	CI	*p*
(Intercept)	0.04	−0.02 – −0.10	0.192
Condition [Treatment]	−0.01	−0.02 – −0.01	0.382
Gender [2]	−0.00	−0.02 – −0.01	0.719
Gender [3]	0.00	−0.03 – −0.03	0.966
Grad [1]	−0.00	−0.02 – −0.01	0.679
Age25-34	0.00	−0.01 – −0.02	0.603
Age35-44	0.00	−0.02 – −0.03	0.779
Age45-54	−0.01	−0.06 – −0.03	0.628
Age [55 and over]	−0.02	−0.12 – −0.09	0.746
Pol_interest_1	−0.01	−0.01 – −0.00	0.059
LR_Score	−0.00	−0.01 – −0.01	0.434
FromMP [1]	−0.00	−0.02 – −0.02	0.935
Lib_Auth	−0.00	−0.02 – −0.01	0.743
WAUse_1	−0.00	−0.01 – −0.00	0.314
News.checking_1	0.01	−0.00 – −0.02	0.102
Social.checking_1	−0.00	−0.01 – −0.00	0.555
Observations		662	
R^2 / R^2 adjusted		0.012 / -0.011	

Table S36 ANCOVA on Post-Treatment scores of reliability assessments (of manipulative items) – data filtered for treatment participants that inputted the game password correctly

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.17	1	0.17	23.97	$p < 0.001$		
F_Rel_Pre	4.18	1	4.18	586.29	$p < 0.001$	.56	[.52, .60]
Condition	0.00	1	0.00	0.10	.752	.00	[.00, .01]
Error	3.25	456	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S37 ANCOVA on Post-Treatment scores of confidence in assessments (of manipulative items) – data filtered for treatment participants that inputted the game password correctly

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.55	1	0.55	70.67	$p < 0.001$		
F_Conf_Pre	3.54	1	3.54	458.50	$p < 0.001$	.50	[.45, .54]
Condition	0.00	1	0.00	0.25	.615	.00	[.00, .01]
Error	3.52	456	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S38 ANCOVA on Post-Treatment scores of sharing measure (of manipulative items) – data filtered for treatment participants that inputted the game password correctly

Predictor	SumofSquares	*df*	MeanSquare	*F*	*p*	$_{partial}\ \eta^2$	$_{partial}\ \eta^2$ 90%*CI* [LL, UL]
(Intercept)	0.32	1	0.32	42.52	$p < 0.001$		
F_Share_Pre	4.85	1	4.85	652.45	$p < 0.001$	.59	[.54, .63]
Condition	0.00	1	0.00	0.67	.413	.00	[.00, .01]
Error	3.39	456	0.01				

Note. LL and UL represent the lower-limit and upper-limit of the partial η^2 confidence interval, respectively.

Table S39 Proportion of rural population across participants' states

State	*n*	*Rural population (%)*	*Weighted Rural*[1]
Bihar	19	89	1685
Chhattisgarh	42	77	3226
Delhi	3	2	8
Haryana	5	65	325
Jharkhand	26	76	1976
Madhya Pradesh	471	72	34100
Rajasthan	120	75	9012
Unknown*	6	69	413
Uttar Pradesh	33	78	2564
		Weighted Mean	73.5

**For missing values, rural proportion of India's national population was imputed*
*[1] Weighted Rural = n * Rural population (%)*
All rural population (%) values sourced from:

Table S40 Distribution between conditions by state

Condition	*State*	*n*
Control	Bihar	15
Treatment	Bihar	4
Control	Chhattisgarh	22
Treatment	Chhattisgarh	20
Control	Delhi	1
Treatment	Delhi	2
Control	Haryana	5
Control	Jharkhand	10
Treatment	Jharkhand	16
Control	Madhya Pradesh	232
Treatment	Madhya Pradesh	239
Control	Rajasthan	62
Treatment	Rajasthan	58
Control	Unknown	4
Treatment	Unknown	2
Control	Uttar Pradesh	14
Treatment	Uttar Pradesh	19

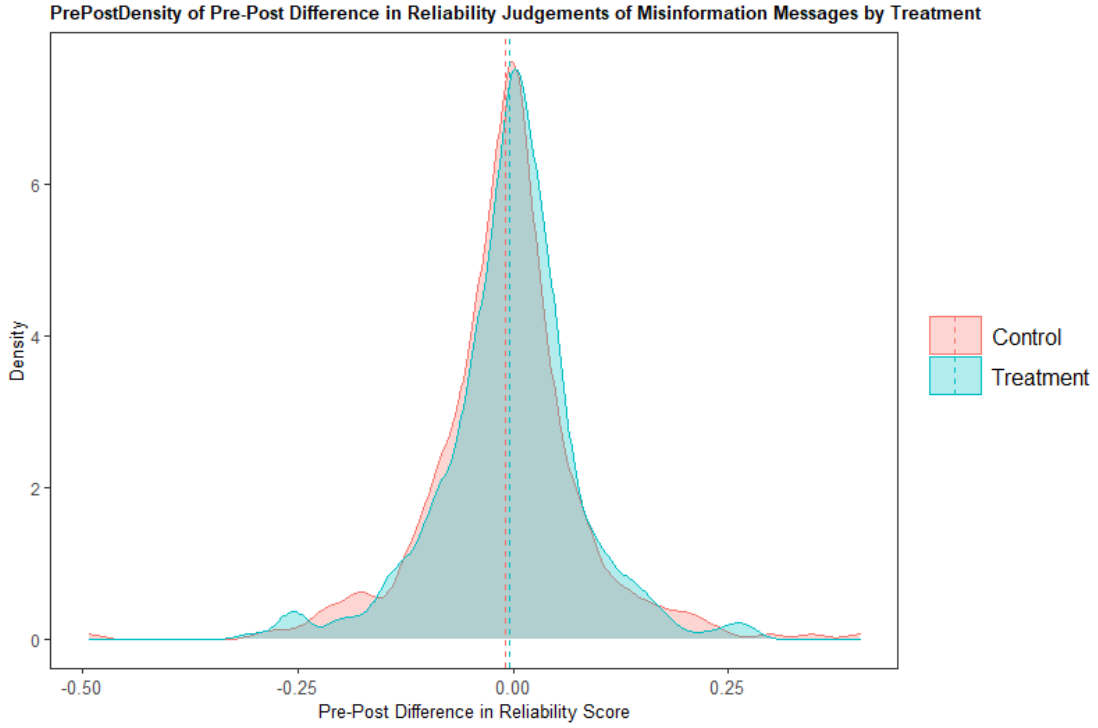


Figure S1 Distribution of Pre-Post Differences in Reliability Judgements of Manipulative Items by Condition

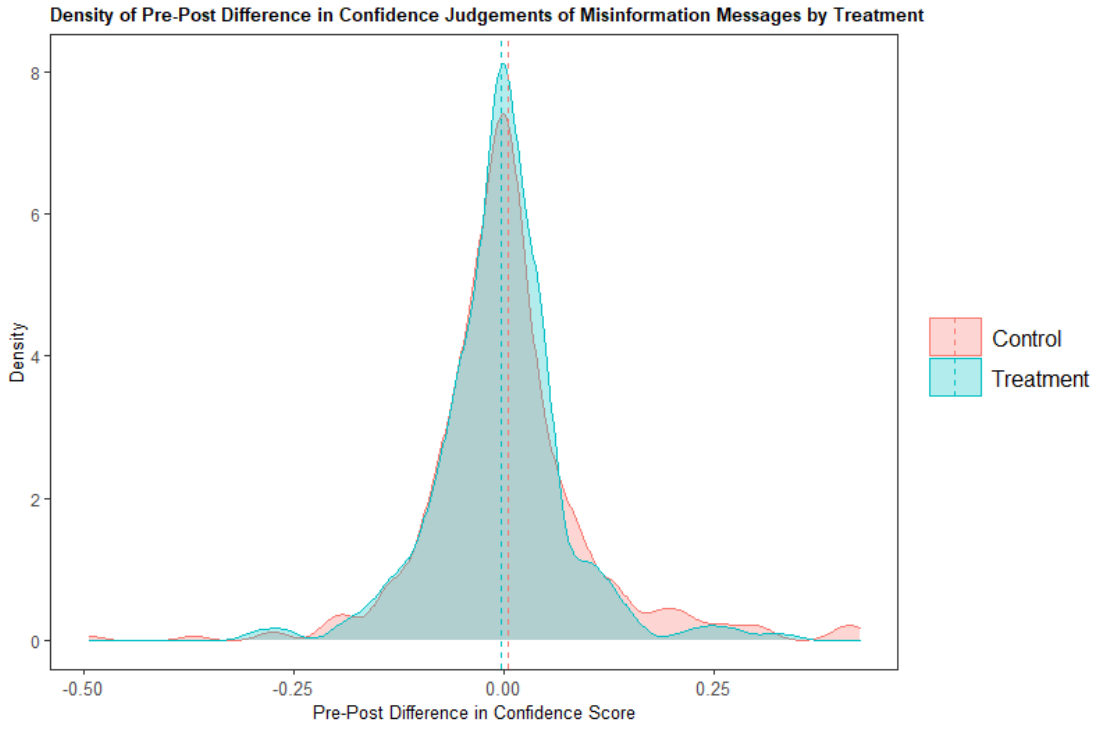


Figure S2 Distribution of Pre-Post Differences in Confidence in Judgements of Manipulative Items by Condition

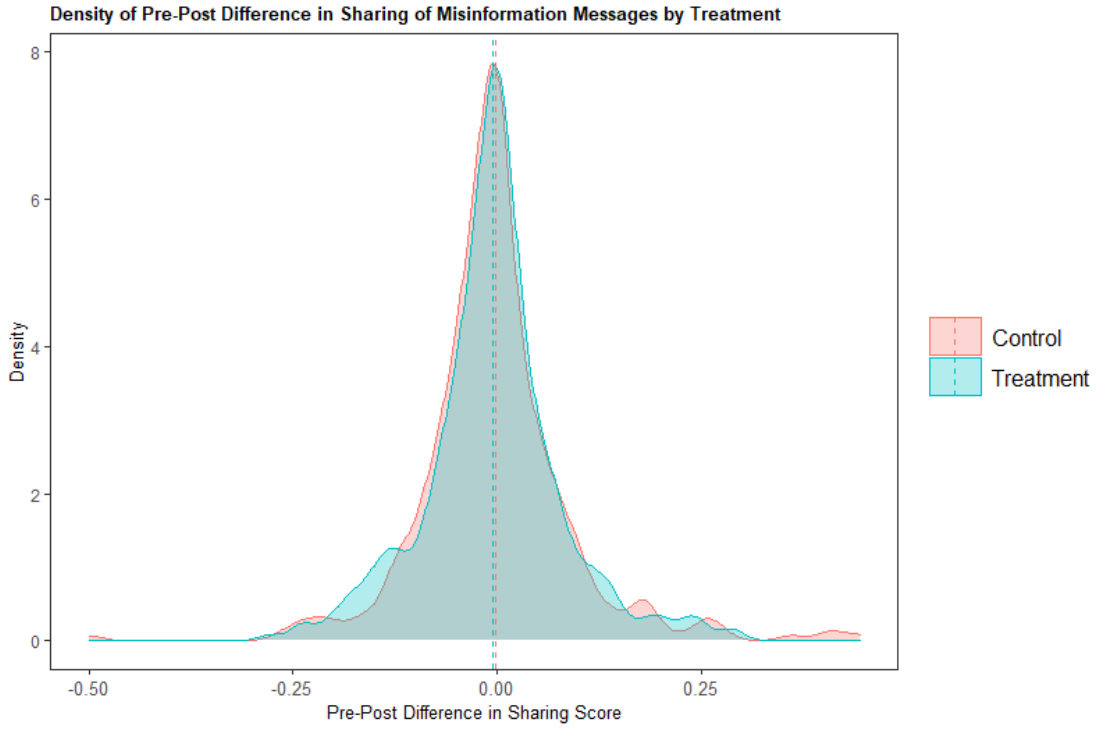


Figure S3 Distribution of Pre-Post Differences in Likelihood to Share Manipulative Items by Condition

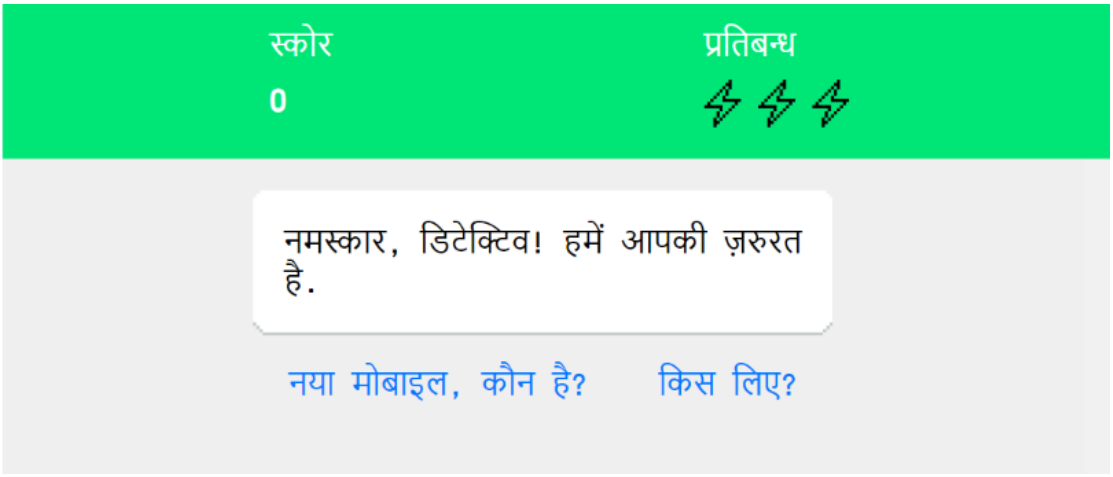


Figure S4 In-game Screenshot - First screen shown after starting the game, introducing the character and motive
Translation:
Green Bar (Left to Right): "Score" "Sanctions"
White Box: "Hello, Detective! We need you"

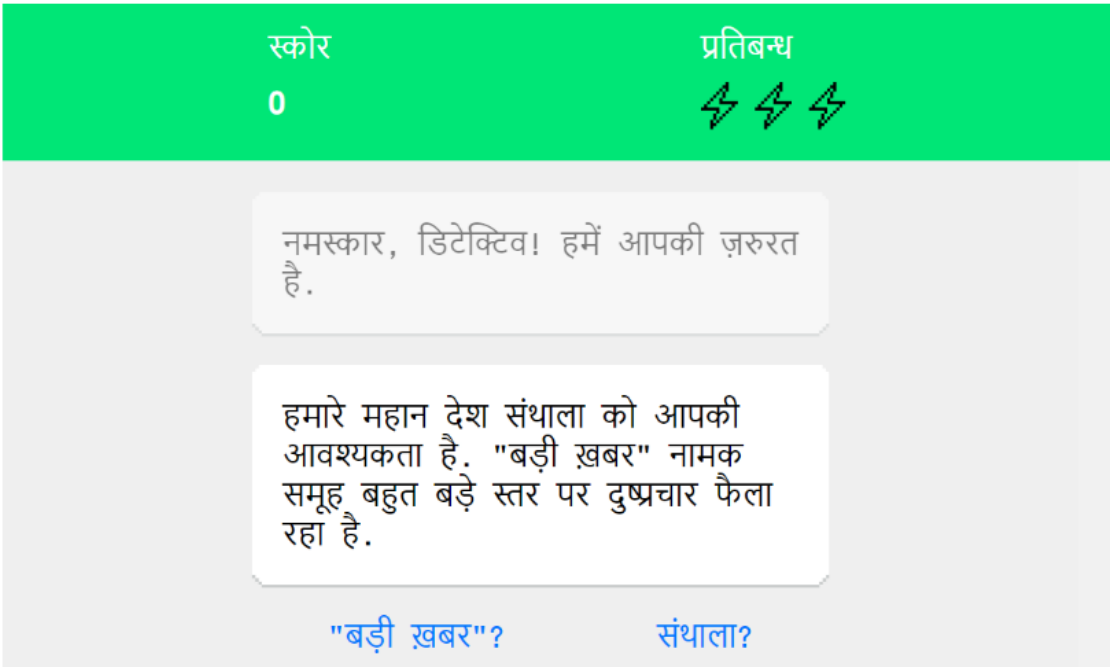


Figure S5 In-game Screenshot - Second screen shown after starting the game, depicting an explanation of the propaganda spreading on WhatsApp.
Translation:
Green Bar (Left to Right): "Score" "Sanctions"
White Box: "Our great country Santhala needs you. A group called "Big News" is spreading propaganda at a very large scale"
Blue Text (Left to Right): "New mobile, who's this?" "For what?"
Blue text: "Big News?" "Santhala?"

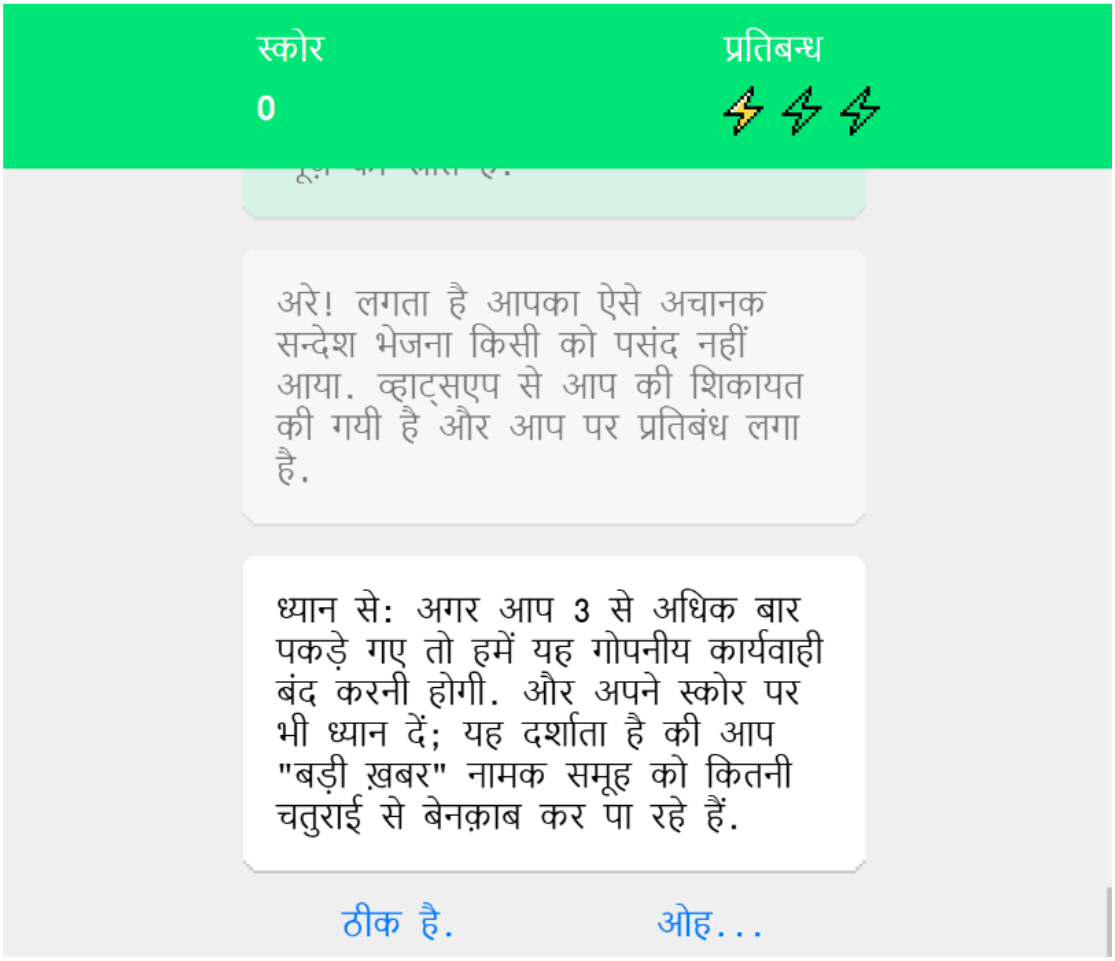


Figure S6 In-game Screenshot - An in-game screenshot explaining the rules of the game.
Translation:
Green Bar (Left to Right): "Score" "Sanctions"
White Box: "Be careful: If you get caught more than 3 times then we have to stop this secrecy. And watch your score as well; this will tell you how much you are exposing the "Bad News" group."
Blue text: "That's fine""Okay..."

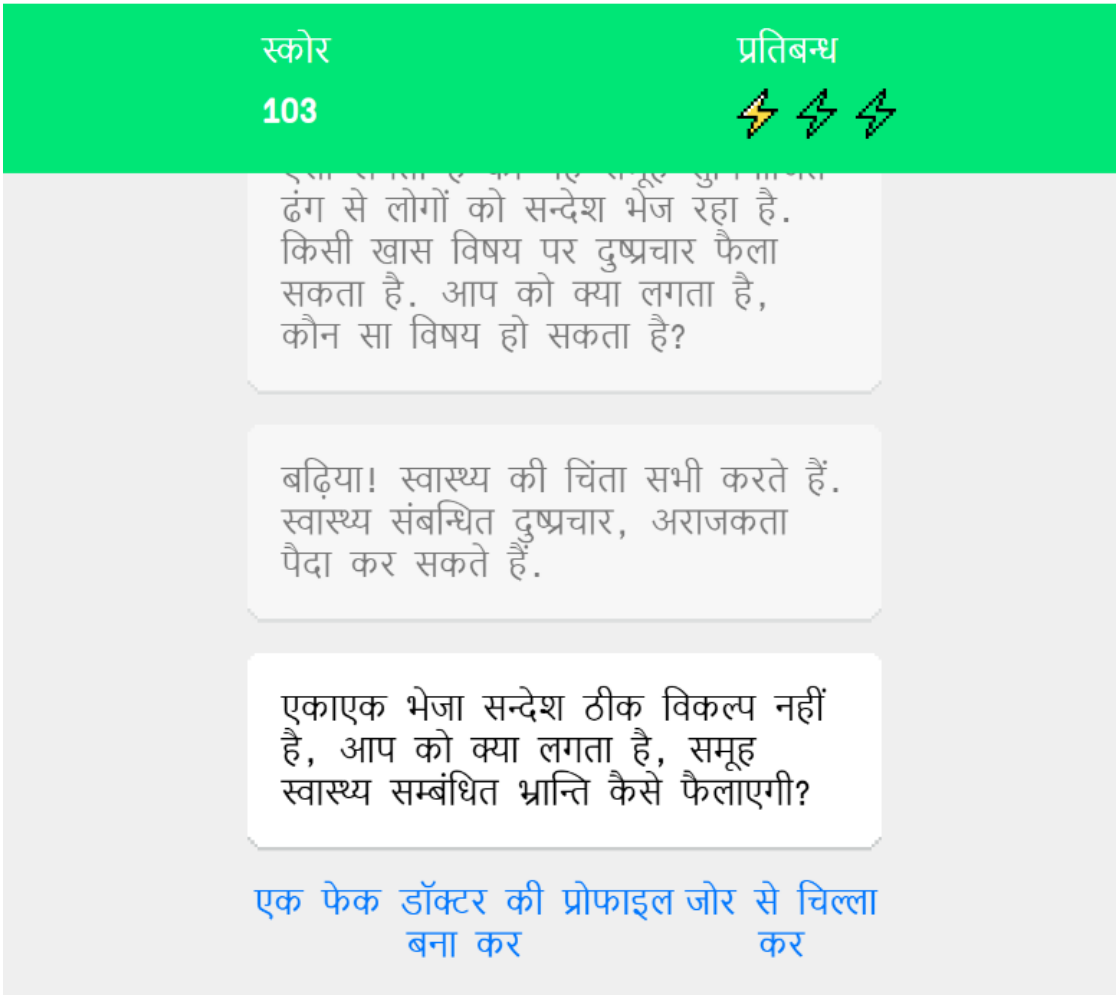


Figure S7 In-game screenshot - Showing how a Fake News technique (using a fake expert) is taught.
Translation: *Green Bar (Left to Right): "Score" "Sanctions"*
White Box: "Just sending a message all of a sudden isn't the right way, what do you think, how will the group spread this health-related misconception?"
Blue text: "By creating a fake doctor profile" "By shouting loudly"

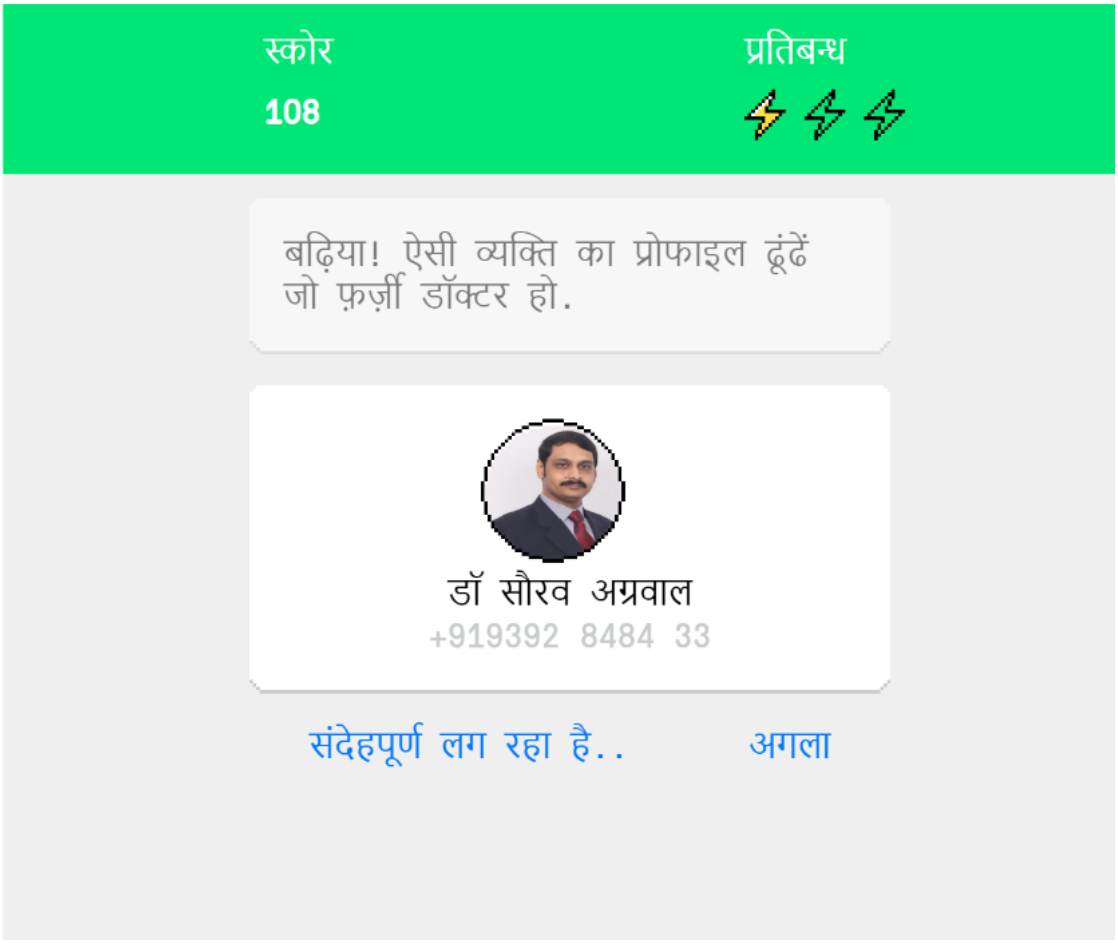


Figure S8 In-game screenshot showing how the Fake News techniques is taught. Continuation of Figure S7. Translation:
Green Bar (Left to Right): "Score" "Sanctions"
Grey Box: "Well done! Find the profile of a person who is a fake doctor"
White Box: "Dr Saurav Agrawal"
Blue Text: "It looks suspicious..." "Next"

Figure S9 Example of a translated manipulative WhatsApp prompt (with English version from another study) intended to show the use of a fake expert.
Screenshot reads: "Hello!
Nowadays it's been very dry.
Even in the rainy season, it does not rain", "Not sure what's happening with the weather these days. Maybe this is happening because of the climate change in the environment",
"Do you think this is happening because of climate change?",
"I'm not sure, it's difficult to say, farming has become very difficult",
"Hello, I am a scientist, climate change is a big reason for whatever is happening in our environment. We have to save our environment.",
"Right, interesting".

Driven to Snack: Simulated Driving Increases Subsequent Consumption

Floor van Meer[1,2*], Stephen Lee Murphy[3], Wilhelm Hofmann[3], Henk van Steenbergen[1,2], Lotte F. Van Dillen[1,2,4]

[1]Institute of Psychology, Leiden University, The Netherlands

[2]Leiden Institute for Brain and Cognition, Leiden University, The Netherlands

[3]Ruhr University Bochum, Germany

[4]Knowledge Centre Psychology and Economic Behaviour, Leiden University, The Netherlands

Received
November 12, 2022
Accepted
March 21, 2023
Published
April 24, 2023
Issued
December 18, 2023

Correspondence
Social, Economic and Organizational Psychology Unit, Leiden University, P.O. Box 9555, 2300 RB Leiden, the Netherlands
a.f.van.meer@leidenuniv.nl

Funding
This research was supported by an Open Research Area grant (Dutch Research Council Grant No. 464-18-105; German Research Foundation Grant HO 4175/7-1).

When individuals eat while distracted, they may compensate by consuming more afterwards. Here, we examined the effect of eating while driving, and explored potential underlying mechanisms. Participants (N = 116, 73.3% female) were randomly allocated to complete a driving simulation (distraction condition) or to watch someone else drive (control condition) while consuming 10g (50.8 kcal) of potato chips. Afterwards, participants rated the taste intensity and hedonic experience, reported stress levels, and were then given the opportunity to eat more chips. As hypothesized, participants consumed more chips after the driving simulation. Stress levels were higher in the driving compared to control condition, but were inversely related to consumption amount, ruling out stress as explanatory mechanism. Saltiness ratings differed between the driving and passive viewing condition, only when controlling for stress. The current findings converge with earlier work showing that distracted eating can drive overconsumption, which in turn can lead to long-term health implications. Limitations, implications, and potential directions are discussed.

Keywords *distracted eating, distraction, food intake, taste perception, consumption*

When eating or drinking, people are frequently exposed to situational stimuli likely to distract them from the sensory experience of consumption. Pre-packaged foods and drinks are increasingly popular and are often consumed while people engage in other activities such as listening to music, using their mobile devices, or commuting. Especially when consumption takes place under cognitively taxing conditions, such as while driving, behind a computer, or while looking at one's smartphone, this practice is likely to increase the amount consumed. For example, children and adults that watch television during their meals have been found to consume more food during (Blass et al., **2006**; Crespo et al., **2001**; Dubois et al., **2008**), and following (Higgs & Woodward, **2009**) the consumption occasion. Likewise, eating while driving, and while listening to loud music, has been found to be less effective in reducing people's desire to eat (Ogden et al., **2016**), and to promote overconsumption (Spence & Shankar, **2010**; Stafford & Dodd, **2013**; Stroebele & de Castro, **2006**). Finally, a meta-analysis examining the effect of distraction during consumption on the amount of food consumed revealed a positive association between these factors (however, one study included in this analysis may have biased the overall effect size Robinson et al., **2013**).

Several mechanisms have been proposed to explain why distracted eating promotes overconsumption, such as reduced awareness of the amount consumed and reduced memory of food intake (Robinson et al., **2013**; Oldham-Cooper et al., **2011**) and compensatory responses to stress (Reichenberger et al., **2018**; Torres & Nowson, **2007**). There is also growing evidence to suggest that the positive link between distraction and consumption amount may be explained by reduced taste perception. For instance, a number of experiments have demonstrated that distraction reduces the taste or odor intensity of sweet, sour, and

Take-home Message

In this study, people consumed more potato chips after eating chips while completing a driving simulation than in a control condition. We had hypothesized that this was due to lowered perceived taste intensity of the potato chips eaten while distracted, but this was only the case when we controlled for stress. Differences in perceived stress did not explain the differences in subsequent consumption amount between the conditions.

bitter solutions, and salty snacks (Hoffmann-Hensel et al., **2017**; Liang et al., **2018**; van der Wal & van Dillen, **2013**), and even promotes increased consumption (Morris et al., **2020**; van der Wal & van Dillen, **2013**). Participants who were distracted by a working memory task while preparing lemonade at their preferred concentration opted for greater amounts of syrup and consumed more salty buttered crackers than participants who experienced minimal distraction (van der Wal & van Dillen, **2013**). Additionally, compared to mildly distracted participants, highly distracted participants exhibited reduced neural taste processing during tasting, while they consumed more during a subsequent ad libitum food test (Duif et al., **2020**). More generally, several recent studies have pointed to the importance of sensory perception, in particular taste intensity, for expectations of fullness and later portion selection (as reviewed in Forde, **2018**). Furthermore, salt intensity predicted ad libitum intake, even when the foods were equally liked (Bolhuis et al., **2012**). However, to our knowledge, no studies have examined both the effect of distraction on perceived taste intensity and palatability of the food consumed *and* how this influences later consumption. Furthermore, previous studies on distracted tasting have used distractions that were either not very ecologically valid (e.g. working memory task van der Wal & van Dillen, **2013**; Duif et al., **2020**; Liang et al., **2018**) or not very cognitively demanding (e.g. listening to music, Stroebele & de Castro, **2006**). Accordingly, the aim of the present study is to investigate the proposed effect using a more ecologically valid distractor – to examine whether eating while driving promotes increased consumption, and whether this effect is explained by reduced taste intensity.

Increased stress levels may provide an alternative explanation for the effect of distracted consumption on increased consumption. That is, it is plausible that driving may imbue stress (Antoun et al., **2017**). For example, participants completing a driving simulation were more stressed when driving themselves than when the simulation was of a self-driving car, evidenced by a higher skin potential response and heart rate (Zontone et al., **2020**). Elevated stress levels have been linked to both increased and reduced food intake (Reichenberger et al., **2018**; Torres & Nowson, **2007**). For instance, ego threat leads to increased snack intake in one study (Wallis & Hetherington, **2004**) but lower snack intake in another (Wallis & Hetherington, **2009**), depending on the type of snacks offered and restrained and emotional eating style. Another factor that may influence whether stress has a positive or negative effect on food intake may be the severity of the stress (Torres & Nowson, **2007**). Thus, we additionally examined the potential role of stress in compensatory consumption following distracted eating (snacking while driving).

Societal and technological developments have increased the frequency in which foods (particularly high-calorie snacks Hirschberg et al., **2016**) are consumed while driving (FoodShopper Monitor, **2018**; Stutts et al., **2005**), thus making this an ideal and realistic scenario in which to test this effect. Furthermore, although multiple studies have found that eating while driving negatively influences driving performance (Dingus et al., **2016**; Irwin et al., **2014**; Young et al., **2008**), the reverse question of whether driving influences eating has so far not been addressed.

The driving context was chosen to be demanding so as to require attention (rather than just routine), and to be representative of everyday demanding driving contexts (e.g., driving on an unfamiliar road, or city traffic during rush hour). We expected that driving would thus induce stress, and mental load. As a result of this higher demand, we hypothesized that driving, relative to control (passive viewing), decreases the perceived taste intensity of salty potato chips. At the same time we expected

that it would lead to greater chip consumption afterwards. As noted, we were less certain about the role of stress in this mechanism, as previous research has observed both increased and decreased consumption following stress. Therefore, we examined the possibility of both a positive and a negative relationship between stress and subsequent consumption. Furthermore, we also explored the effects of distraction on the hedonic aspects of taste. We hypothesized that distraction decreases perceived taste intensity but may not affect hedonic ratings, as the hedonic value of consumption varies greatly between individuals but is stable within individuals and as this has not been consistently linked with actual consumption (DiFeliceantonio et al., **2018**; McCrickerd & Forde, **2015**; Tang et al., **2014**). Therefore, we did not think it likely that the hypothesized effect of distraction on consumption could be explained by changes in hedonic ratings. Moreover, we explored whether participants' driving experience was a potential moderator of our proposed effects of distraction on taste perception and consumption since this may affect how demanding and stressful the driving manipulation was for each participant. Finally, since some previous studies have found that the effects of distraction on consumption vary with individual differences in restrained eating (Boon et al., **2002**; Ogden et al., **2016**), this was included as a control variable.

Methods

Participants and Design

One hundred nineteen English speaking Leiden University students in possession of a driver's license (car) participated in exchange for course credit or money (€3.50) and were randomly assigned to a simulated driving or control condition. Smoking or having allergies were exclusion criteria. Participants were requested not to eat and to only drink water two hours prior to the start of the study. Of the sample, three cases were excluded because they fell outside the proposed age range of 18-30 years (ages 45 and 60 years, > 3 *SD*s from the mean; for one participant age was not known). An additional three participants initiated but did not complete the study and were therefore also excluded from further analyses. Repeating the analyses including these participants did not change the results. The remaining sample for analyses thus consisted of 116 participants (30 men, mean age 22.30 years, *SD* = 4.98 years) evenly distributed over the two conditions (n = 58 each). The two groups did not differ on the number of men and women, age, nationality, or total Restrained Eating Score (see Supplemental Table 4).

The main dependent variables were taste intensity of the potato chips and the amount of calories consumed. In addition, stress levels and hedonic ratings were considered. Individual differences in driving experience and restrained eating were examined as potential moderators. The research questions and procedure were approved by the ethical committee of the Leiden University Psychology Institute (CEP19-0301/146). All procedures performed were in accordance with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants involved in the study. The overall design, research question and hypotheses were specified in the ethics proposal prior to the start of data collection. The ethics proposal, raw data and analysis script can be obtained from: osf.io/twg9r/.

Procedure

Before engaging in the experiment, participants were seated behind a desk with a laptop on which a short introductory text was displayed that informed them that the study was about multitasking while driving. After providing informed consent, participants next reported their driving experience. Following this, they were randomly assigned to the driving or control condition (see *Driving simulation* for details), and asked to sit in the driver's seat of the simulator where they were provided instructions for the driving simulation. Participants were then provided with a bowl of potato chips (10 grams, 50.8 kcal) and instructed to consume them all during the driving simulation. The chips used were Lays Classic salted potato chips. All participants consumed the entire 10 grams/50.8 kcal. Participants then completed the driving simulation. Following

Figure 1 The set-up of the driving simulator used in both the experimental driving and passive viewing control conditions. It consisted of a chair, steering wheel, pedals and a 23-inch flat screen. A PlayStation 3 and the game Gran Turismo (Yamauchi, **2013**) were used to simulate the actual driving experience. Participants drove (or viewed a recorded video of) three laps on the Twin Motegi Course.

Materials

Driving Simulation

To create a realistic and demanding driving context, a set-up was built that consisted of a chair, steering wheel, pedals, and a 23-inch flat screen (see Fig.1). A PlayStation 3 and the game Gran Turismo (Yamauchi, **2013**) were used to simulate a realistic driving experience. Participants were seated in the driving chair and it was explained how they could speed up, break and steer. Participants were asked to drive three laps on the Twin Ring Motegi course that consisted of two straight sections, a large bend and 2 sharp bends. Participants were told they should drive as well as they could. Driving the three laps took three minutes on average. If a participant took longer than 10 minutes to complete the laps the simulation was stopped, however, none of the participants took longer than 10 minutes to drive the three laps. To create a comparable situation in the control condition, the same driving simulator was used. The participants in the control group acted as co-driver/passenger and did not actually drive themselves. Instead, a three-minute recorded video was played, showing the exact same three laps of the Twin Ring Motegi course that the participants drove in the experimental driving condition.

Driving Experience

Three questions addressed participants' driving experience: 'How many years do you have your driver's license?', 'How often do you drive on average per week?' and 'How many kilometers did you cover on average in the last year?'. The three items were answered on five-point Likert scales. These included respectively, driving years ranging from 1 (up to 1 year), increasing with each scale point with 1 year to a maximum of 5 (over 7 years); driving frequency ranging from 1 (once a week) increasing with each scale point with one time per week to a maximum of 5 (7 times per week); and driving distance ranging from 1 (1,000km per year) increasing with each scale point with 1,000km per year to a maximum of 5 (7,000 km per year).

the driving simulation, participants returned to the desk to report their ratings, stress levels, age, sex, and ethnicity on the laptop. Participants were then instructed to wait in the room while the experimenter collected debriefing forms from the adjacent room (wait time held constant at three minutes), and that if they wanted, they could eat the rest of the potato chips (the remaining 15 g (76.2 kcal) from the 25 g party bag, in a bowl on the same desk). The Netherlands Nutrition Centre states 30 g as the average portion size of chips in the Netherlands (Voedingscentrum, **n.d.**). Participants were told that these potato chips were left over from the party bag and that they were free to consume them all. Finally, all participants were debriefed, thanked, and compensated for their participation.

Taste Intensity

Participants rated the potato chips on three items relating to taste intensity, namely 'saltiness', 'sourness', and 'sweetness', on seven-point Likert scales ranging from 1 (*not at all*) to 7 (*very*). Sweetness and sourness ratings were included as catch trials, to establish that participants were not merely guessing when assessing the potato chips' flavor. Furthermore, the sweetness and sourness ratings serve as a baseline measure since we do not expect them to differ between conditions.

Hedonic Rating

Participants next rated the potato chips on three more items relating to hedonic experience, namely 'quality', 'tastiness', and 'crunchiness', on the same seven-point Likert scales ranging from 1 (*not at all*) to 7 (*very*).

Stress levels

Participants were asked five questions pertaining to their experiences of stress during the simulation: 'How relaxed were you during the driving simulation?' (reversed), 'How much did you have the feeling that you were in control during the driving simulation?' (reversed), 'How rushed did you feel during the driving simulation?', 'How nervous were you during the driving simulation?', and 'How well did you perform during the driving simulation?' (reversed). All questions were answered on a six-point Likert scale ranging from 1 (*not at all*) to 6 (*very*).

Calories Consumed

The number of calories consumed was determined by weighing the bowl with the remaining chips once the participant had left and subtracting this from its initial weight. The weight in gram was then multiplied by the amount of kcal/g (5.08).

Data Preparation

All data preparation steps and statistical analyses were performed in R (R Core Team, **2019**) and can be retrieved from the OSF page: osf.io/twg9r/. Distribution of the variables was examined by visual inspection, Shapiro-Wilks test and Levene's test. Since some of the variables were skewed, robust regression using the rlm function of the R package MASS was used throughout for consistency.

Robust regression was used to examine the differences between conditions unless otherwise specified (see Results). For each dependent variable (taste intensity, hedonic rating and number of calories consumed) we first estimated a full factorial model that included main effects and interaction of the experimental factor (Driving, Control) and Driving Experience. If the interaction term was not statistically significant, subsequently models with only the main effects of the experimental factor and Driving Experience were estimated. Afterwards, we calculated the Bayes Information Criterion (BIC) in order to see which model performed best.

Reliability analysis revealed saltiness was poorly associated with sourness and sweetness (Cronbach's $\alpha = 0.31$), as expected, and so these ratings were therefore examined separately. The items assessing hedonic rating and stress showed adequate reliability (Cronbach's alphas of respectively .69 and .79) and were therefore averaged into two overall scores. The three items that assessed driving experience were only moderately associated (Cronbach's $\alpha = 0.54$), but driving distance correlated significantly with both frequency ($r = 0.48$) and years of license ($r = 0.34$), with the latter two being uncorrelated ($r = 0.06$). Even though each item thus seemed to tap into a somewhat different aspect of driving experience, they were nevertheless averaged to form a broad index of driving experience.

Subsequent t-test analyses confirmed that driving experience in years, frequency and distance did not vary across conditions (ts < 1.42, ps > 0.153, so that these could be incorporated as moderator variables into the regression models for taste ratings and consumption. Table 1 depicts the raw means and standard deviations of the three Driving Experience items (years, frequency and distance) as a function of condition. Since driving experience was highly skewed towards the lower end (see Figure 3a in the Supplement section), quartile scores were used in the analyses.

Control analyses showed that men had more driving experience ($M = 2.10$, $SD = 0.71$) than women ($M = 1.71$, $SD = 0.76$; $F(1,114) = 6.30$, $p = 0.01$). Moreover, men consumed more

Table 1 Means and standard deviations of driving experience (in years, distance and frequency) as a function of condition (driving; control).

Condition	Driving Years[1]	Driving Distance	Driving Frequency
Driving	2.13 (1.23)	1.90 (1.28)	1.45 (.81)
Control	2.39 (1.17)	1.63 (1.02)	1.53 (.86)

[1] *The driving experience items were answered on five-point Likert scales ranging from respectively 1 (up to 1 year/once a week /1000km per year) to 5 (over 7 years/ 7 times per week /7000 km per year).*

calories than women irrespective of the experimental condition (men: M = 100.0 kcal, SD = 27.9; women: M = 71.7 kcal, SD = 25.0; b = -9.09, SE = 2.04, $t(111)$ = -4.45, p > .001). Therefore, we corrected for gender in all our analyses.

Results

Effects of Driving on Calories consumed

Table 2 and Figure 2 depict the mean and standard deviation/error for potato chips consumed in kcal during the driving manipulation and the follow-up free consumption test as a function of condition (Driving; Control).

Inspection of the histograms revealed that the number of calories consumed was not normally distributed, but had a bimodal distribution with many observations at the scale extremes (50.8, 127.0; see supplemental Fig. S6.b for histograms per condition). More specifically, during the free consumption period 52% of participants consumed no chips and 20% of participants consumed all of the chips. Given how many participants consumed the maximum amount of chips available, it is likely that the mean consumption amount would have been higher if it had not been restricted (i.e., censoring effect is likely). Therefore, we applied Tobit regression analyses[1] (i.e., censored regression models Tobin, **1958**), using the R package censReg (Henningsen, **2010**).

This Tobit regression analysis with calories consumed as dependent variable and main effects and interaction of the experimental factor (Driving, Control) and Driving Experience showed no significant interaction term, so a model with only main effects was estimated. There was a significant main effect of the driving manipulation, b = -19.43, p = 0.026. As hypothesized, participants consumed more potato chips when driving (M = 84.3 kcal, SE = 4.11), compared to passively watching the same route (M = 72.9 kcal, SE = 3.24). Driving Experience did not have a main effect. The BIC for the model with only main effects was lower than the model with the interaction term (BIC 3.15).

[1]We also analyzed the number of calories consumed using robust regression models. These yielded comparable results, see: osf.io/twg9r/.

Effects of Driving on Taste Intensity

Table 2 and Figure 2 depict the means and standard deviations for all taste intensity ratings as a function of condition (Driving; Control).

Robust regression analyses incorporating main effects and interaction of the experimental factor (Driving, Control) and Driving Experience were conducted to examine the effects on saltiness ratings. The BIC for the model with only main effects was lower than the model with the interaction term (BIC 2.67). Contrary to our first hypothesis, we did not observe a significant main effect of condition, b = 0.42, SE = .23, $t(111)$ = 1.84, p = 0.067. As predicted, participants rated the potato chips as less salty when they were driving themselves (M = 4.43, SE = 0.16), than when they were attending a recording of the same route being driven by someone else (M = 4.74, SE = 0.15), but this difference did not reach the threshold for significance. Control analyses confirmed that the driving manipulation likewise did not significantly impact participants' sourness and sweetness ratings (ts < 0.3, ps > .54) with very similar intensity ratings across conditions (see Fig.2). The potato chips were generally perceived to be minimally sweet (M = 2.01, SD = 1.15) and sour (M = 1.72, SD = 0.96). Taken together, even though the intensity ratings showed the expected pattern, we found no robust proof that driving interfered with participants' processing of the saltiness of the chips.

There was no main effect or interaction effect of Driving Experience.

Table 2 Means and standard deviations of the various taste ratings (1 - not at all to 7 - very) and amount consumed in kcal as a function of condition (driving; control).

Condition	Salty	Sweet	Sour	Quality	Tasty	Crunchy	Total Calories Consumed
Driving	4.43 (1.19)	2.00 (1.27) 1.79 (1.04)	1.79 (1.04)	4.85 (1.14)	5.17 (1.35)	5.12 (1.20)	84.3 (31.1)
Control	4.74 (1.15)	2.05 (1.03)	1.69 (0.90)	4.67 (1.37)	4.98 (1.40)	5.22 (1.13)	72.9 (24.7)

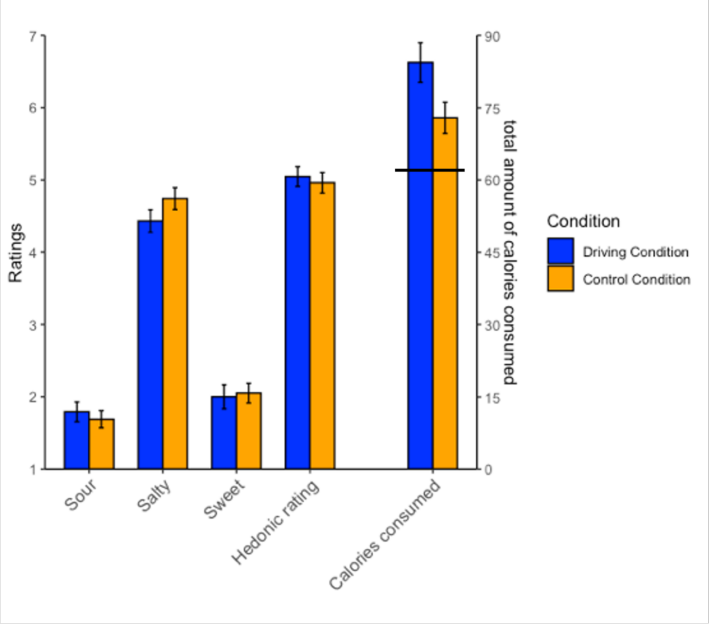


Figure 2 Mean of taste intensity ratings, hedonic ratings and total calories of chips consumed per condition. "Hedonic rating" here reflects the mean of the "quality", "crunchy" and "tasty" ratings. Total amount of calories includes the standard amount of 50.8 kcal of chips eaten during the manipulation, as indicated by the horizontal black line. Error bars reflect standard error.

Explorations of Hedonic Rating and Driving-induced Stress as Alternative Explanation

We also explored whether driving altered hedonic aspects of the consumption experience. A robust regression model with hedonic rating as dependent variable and main effects and interaction of the experimental factor (Driving, Control) and Driving Experience was estimated. Table 2 depicts the means and standard deviations for the three hedonic ratings as a function of condition (Driving; Control). These showed that participants rated hedonic aspects no different in the driving condition (M = 5.08, SE = 0.41) than the control condition (M = 4.95, SE = 0.42, b = -0.18, p = 0.782). When the items were analyzed separately, this did not yield any significant differences either (ps>0.356). Finally, driving experience did not significantly impact hedonic rating (p=0.344) nor was there a significant interaction between Driving Experience and condition on hedonic rating (p = 0.125).

We next examined whether the effects of driving on perception and consumption resulted from driving-induced stress as opposed to distraction. Table 3 depicts the means and standard deviations for the five stress ratings as a function of condition (Driving; Control). These revealed that all five items were affected by the driving manipulation; participants were significantly less relaxed, and felt significantly more in control, rushed, nervous, and performing well while driving than while in the passive viewing condition, ts>2.93, ps<0.020. This confirms that driving compared to passive viewing heightened participants' stress levels. There was no interaction between Driving Experience and the driving manipulation on perceived stress levels (p = 0.17).

To test whether the effect of condition on the amount of food consumed could be explained by the difference in experienced stress, a Tobit regression analysis was performed with calories consumed as the dependent variable and the main effects and interactions of the experimental factor (Driving, Control), stress, and Driving Experience. Since the full factorial did not show a significant effect of the interaction term with Driving Experience, subsequently a model was estimated with calories consumed as the dependent variable and

Table 3 Means and standard deviations (between brackets) of the various stress ratings (1- not at all to 7 – very) as a function of condition (driving; control).

Condition	Relaxed	In control	Rushed	Nervous	Performed well
Driving	3.28 (1.32)	3.20 (1.30)	3.82 (1.24)	3.17 (1.45)	3.95 (.95)
Control	4.44 (1.34)	1.56 (.88)	2.88 (1.35)	2.56 (1.41)	2.63 (1.07)

the main effects and interactions of the experimental factor (Driving, Control) and stress and only a main effect for Driving Experience (BIC 9.43). The analysis showed a main effect for both conditions, b = -25.82, SE = 32.14, t =-2.60, p = 0.009, and stress, b = -15.83, SE = 7.60, t = -2.08, p = 0.037. Interestingly, the effect of stress on consumption was negative, which means that participants who felt more stressed ate less. This suggests that increases in stress did not explain increased consumption following driving. Additionally, there was a significant interaction effect of driving manipulation and stress on calories consumed: b = 18.22, SE = 9.29, t = 1.96, p = 0.038. Although there was no significant main effect of stress on consumption when the analyses were done in the respective conditions, a Fischer r to z comparison confirmed that the slopes of the effect of stress on calories consumed in the driving and control condition were different, Z = 2.00, p = 0.05. In the driving condition stress had a stronger negative effect on calories consumed (r = -0.37) compared to the control condition (r = -0.12). There was no significant main effect of stress on consumption when the analyses were done in the respective conditions. When added as covariate to the overall regression model, stress did not explain the main effect of driving on calories consumed.

In conclusion, stress had a negative effect on consumption. So, even though participants felt more stressed in the driving condition, stress did not account for the difference in calories consumed between the driving and control condition.

There was no effect of stress on saltiness ratings or any interactions between stress and condition or driving experience on saltiness ratings (all ps >0.35). However, when stress was entered into the model, the effect of condition on saltiness ratings became significant (b = 0.54, SE = 0.26, t(111) = 2.035, p = 0.0454).[2]

In order to examine the relationships between all factors of interest and to take into account that driving experience and stress are assessed by multiple items, we used structural equation modeling (SEM) to estimate path models using the lavaan package in R (Rosseel, **2012**). Figure 6 in the supplementary section depicts the model that was tested. Calories consumed was the dependent variable, driving condition was included as a predictor and taste perception (saltiness ratings) and stress were assessed as possible mediators between driving condition and calories consumed. Stress and driving experience were modeled as latent variables. Furthermore, gender was added to the model as a control variable. Since lavaan does not support interactions with latent variables, no interactions were modeled. Model fit indices show a poor fit for the model (CFI = 0.69; $SRMR$ = 0.12; $RMSEA$ = 0.14 (90% CI: 0.12 to 0.17). The model showed a significant effect of driving condition on calories consumed (b = 13.62, SE = 6.14, p = 0.027) but did not indicate stress or saltiness ratings as a mediator of this effect via a direct or indirect path (Figure 6; analysis script on OSF).

General Discussion

In this study we aimed to build on previous studies that found that distraction increased

[2] Restrained Eating as measured by the Restrained Eating Scale (Polivy et al., **1978**) was examined as a potential moderator as well. There was no difference in Restrained Eating between the driving (M=13.0; SD=5.82) and control condition (M = 11.9; SD = 4.76). Restrained Eating did not interact with any of the variables of interest. Adding total Restrained Eating score as a covariate did not change the overall pattern of results, see: osf.io/twg9r/.

consumption by examining possible explanations of the effect in a practically relevant setting. To do so, in a simulated driving experiment, we examined whether snacking while driving would result in greater consumption afterwards. We furthermore investigated whether this effect could be explained by reduced taste intensity while driving or by driving-induced stress.

In support of our predictions, participants who engaged in the driving simulation while consuming potato chips, consumed more potato chips during a follow-up free consumption test than participants who merely watched a recording. There were some indications that the driving simulation lowered the saltiness ratings of the chips. Other sensory ratings and hedonic ratings were unaffected by driving.

Many different (complementary) explanations have been proposed for the mechanism which makes people consume more after or during distracted consumption, including reduced memory for food intake or health goals, disrupted influence of satiation, and dishabituation (Forde, **2018**; Robinson et al., **2013**). In the current study, we found some indications that lowered taste perception may be an interesting component to consider when studying the mechanism behind overconsumption after distracted eating.

Our finding that distraction may reduce perceptions of saltiness supports previous literature demonstrating this effect (van der Wal & van Dillen, **2013**; Liang et al., **2018**; Duif et al., **2020**). As taste intensity has been found to correlate negatively with food intake (Forde et al., **2013**), lowered experienced taste intensity during distracted eating may lead to increased food consumption. Furthermore, when distracted, taste information may not be processed in a way that leads to satisfaction or satiation. For example, consuming a high calorie drink under high perceptual load led to lower satiety than when the same drink was consumed under low perceptual load (Morris et al., **2020**). Future studies could examine the effect of distracted consumption on satiation/satiety and how this relates to taste perception and other outcomes.

Perceived stress was examined as an alternative explanation of the effect of distracted consumption on subsequent consumption. Whereas participants reported more stress after driving than after watching someone else drive, self-reported stress yielded an opposite effect on consumption, with participants consuming *fewer* rather than more potato chips. The phenomenon that acute stress can reduce food intake has been attributed to physiologic changes that occur after acute stress and that might be expected to temporarily reduce food intake, e.g., slowed gastric emptying and shifting of blood from the gastrointestinal tract to muscles (Torres & Nowson, **2007**).

Several previous studies have found an effect of restrained eating on the relationship between stress and consumption (Wallis & Hetherington, **2004**; Wallis & Hetherington, **2009**). However, we did not find an effect of restrained eating on consumption or any interaction between restrained eating, stress or driving manipulation. This could possibly be explained by the fact that restrained eating scores in our sample were low.

In conclusion, higher perceived stress was associated with lower consumption of potato chips. Therefore, the finding that the driving manipulation increased intake could be not accounted for by the driving induced stress.

A strength of the current study is the use of a realistic and practically relevant distractor and consumption situation. This study aimed for a control condition that matched the sensory input during driving and thus only differed from the experimental condition in the mental load and stress induced. As a result, the participants in the control condition were probably still somewhat distracted and this might have created a conservative test of our hypotheses.

Original Purpose

In this study, we aimed to examine the effect of eating while driving, and potential underlying mechanisms. We hypothesized that eating while driving would reduce taste perception, which would in turn cause participants to overconsume afterwards to compensate. Based on previous research, we expected that taste perception, but not hedonic preference, would be diminished during distracted consumption. We furthermore wanted to examine the effect of stress experienced during the driving simulation as an alternative explanation.

However, this way, any differences in consumption and perceived taste intensity between the conditions could be attributed to differences in the availability of mental capacity. It is possible that the smaller effect sizes have caused our study to be underpowered to detect the effect of driving condition on saltiness ratings. Future research could examine variations in mental load and stress further by comparing different levels of distraction during consumption, e.g., high distraction, low distraction, no distraction and targeted attention through mindful eating instructions in a larger sample.

The current study also has its limitations. Whereas standardization of the consumption amount during the driving manipulation allowed us to examine differences in compensatory consumption, one limitation of the study is the limited amount that could be consumed later. The mean difference in the amount of potato chips consumed after driving or passively watching was only 11 kcal. However, these 11 kcal were consumed in addition to the 50.8 kcal that participants already ate during the driving distraction or passively viewing. In addition, a substantial proportion of the sample consumed the entire additional 15 grams or 76 kcal, which might indicate that they would have consumed more had they had the opportunity to do so. To further examine the magnitude and practical relevance of the compensatory consumption effect, future studies could examine ad libitum intake following distracted eating.

Although participants were requested not to eat in the two hours prior to the start of the study, subjective hunger was not assessed. However, since participants were randomly assigned to the driving or control condition, possible variability in hunger status is unlikely to have caused the difference in subsequent consumption between conditions.

We did not assess how much experience with playing video games participants had. In addition to driving experience, this may have affected how challenging the driving simulation was for participants.

Lastly, the relatively young age, low driving experience, high education level and unbalanced gender ratio of our sample limits the generalizability of our results. Furthermore, ethnicity was not assessed. Future studies could extend our findings in broader samples that are more representative of the general population.

Conclusion

Using a realistic but lab-controlled driving simulation, the findings reported here provide additional support for the notion that distracting consumption settings may have long-term health implications, through their contribution to overconsumption of unhealthy products. This pushes the need for a better understanding of what these settings look like in people's daily lives and how consumption settings can be changed. The current research provides some preliminary evidence that taste perception, and especially perceived taste intensity, may be a relevant aspect to consider when examining the mechanism through which distracted eating leads to overconsumption.

Funding

This research was supported by an Open Research Area grant (Dutch Research Council Grant No. 464-18-105; German Research Foundation Grant HO 4175/7-1).

Author contributions

LFvD developed the study design and oversaw the data collection; FvM and LFvD performed the primary data analyses and HvS provided additional analyses; FvM, LFvD, HvS, SLM, and WH prepared the manuscript; all authors provided critical revisions and approved the final version of the manuscript.

Acknowledgements

We thank Maureen Botz, Lukas Sutorius and Marit van Wijncoop for their assistance during data collection.

References

Antoun, M., Edwards, K. M., Sweeting, J., & Ding, D. (2017). The acute physiological stress response to driving: A systematic review (J. Xu, Ed.). *PLOS ONE*, *12*(10), e0185517. https://doi.org/10.1371/journal.pone.0185517 (see p. 58).

Blass, E. M., Anderson, D. R., Kirkorian, H. L., Pempek, T. A., Price, I., & Koleini, M. F. (2006). On the road to obesity: Television viewing increases intake of high-density foods. *Physiology & Behavior*, *88*(4-5), 597–604. https://doi.org/10.1016/j.physbeh.2006.05.035 (see p. 57).

Bolhuis, D. P., Lakemond, C. M., de Wijk, R. A., Luning, P. A., & de Graaf, C. (2012). Effect of salt intensity in soup on ad libitum intake and on subsequent food choice. *Appetite*, *58*(1), 48–55. https://doi.org/10.1016/j.appet.2011.09.001 (see p. 58).

Boon, B., Stroebe, W., Schut, H., & Ijntema, R. (2002). Ironic processes in the eating behaviour of restrained eaters. *British Journal of Health Psychology*, *7*(1), 1–10. https://doi.org/10.1348/135910702169303 (see p. 59).

Crespo, C. J., Smit, E., Troiano, R. P., Bartlett, S. J., Macera, C. A., & Andersen, R. E. (2001). Television watching, energy intake, and obesity in US children. *Archives of Pediatrics & Adolescent Medicine*, *155*(3), 360. https://doi.org/10.1001/archpedi.155.3.360 (see p. 57).

DiFeliceantonio, A. G., Coppin, G., Rigoux, L., Thanarajah, S. E., Dagher, A., Tittgemeyer, M., & Small, D. M. (2018). Supra-additive effects of combining fat and carbohydrate on food reward. *Cell Metabolism*, *28*(1), 33–44 (see p. 59).

Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences*, *113*(10), 2636–2641. https://doi.org/10.1073/pnas.1513271113 (see p. 58).

Dubois, L., Farmer, A., Girard, M., & Peterson, K. (2008). Social factors and television use during meals and snacks is associated with higher bmi among pre-school children. *Public Health Nutrition*, *11*(12), 1267–1279. https://doi.org/10.1017/s1368980008002887 (see p. 57).

Duif, I., Wegman, J., Mars, M. M., de Graaf, C., Smeets, P. A., & Aarts, E. (2020). Effects of distraction on taste-related neural processing: A cross-sectional fMRI study. *The American Journal of Clinical Nutrition*, *111*(5), 950–961. https://doi.org/10.1093/ajcn/nqaa032 (see pp. 58, 65).

FoodShopper Monitor. (2018). FSIN FoodShopper Monitor 2019. In *Foodservice instituut nederland*. https://fsin.nl/foodshoppermonitor (see p. 58).

Forde, C. G. (2018). From perception to ingestion: The role of sensory properties in energy selection, eating behaviour and food intake. *Food Quality and Preference*, *66*, 171–177 (see pp. 58, 65).

Forde, C. G., van Kuijk, N., Thaler, T., de Graaf, C., & Martin, N. (2013). Oral processing characteristics of solid savoury meal components, and relationship with food composition, sensory attributes and expected satiation. *Appetite*, *60*, 208–219. https://doi.org/10.1016/j.appet.2012.09.015 (see p. 65).

Henningsen, A. (2010). Estimating censored regression models in R using the censReg Package. *R Package Vignettes*, *5*, 1–12 (see p. 62).

Higgs, S., & Woodward, M. (2009). Television watching during lunch increases afternoon snack intake of young women. *Appetite*, *52*, 39–43 (see p. 57).

Hirschberg, C., Rajko, A., Schumacher, T., & Wrulich, M. (2016). The changing market for food delivery. *Mckinsey and Co*. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-changing-market-for-food-delivery (see p. 58).

Hoffmann-Hensel, S. M., Sijben, R., Rodriguez-Raecke, R., & Freiherr, J. (2017). Cognitive load alters neuronal processing of food odors. *Chemical Senses*, *42*(9), 723–736. https://doi.org/10.1093/chemse/bjx046 (see p. 58).

Irwin, C., Monement, S., & Desbrow, B. (2014). The influence of drinking, texting, and eating on simulated driving performance. *Traffic Injury Prevention*, *16*(2), 116–123. https://doi.org/10.1080/15389588.2014.920953 (see p. 58).

Liang, P., Jiang, J., Ding, Q., Tang, X., & Roy, S. (2018). Memory load influences taste sensitivities. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02533 (see pp. 58, 65).

McCrickerd, K., & Forde, C. G. (2015). Sensory influences on food intake control: Moving beyond palatability. *Obesity Reviews*, *17*(1), 18–29. https://doi.org/10.1111/obr.12340 (see p. 59).

Morris, J., Vi, C. T., Obrist, M., Forster, S., & Yeomans, M. R. (2020). Ingested but not perceived: Response to satiety cues disrupted by perceptual load. *Appetite*, *155*, 104813. https://doi.org/10.1016/j.appet.2020.104813 (see pp. 58, 65).

Ogden, J., Oikonomou, E., & Alemany, G. (2016). Distraction, restrained eating and disinhibition: An experimental study of food intake and the impact of 'eating on the go'. *Journal of Health Psychology*, *22*(1), 39–50. https://doi.org/10.1177/1359105315595119 (see pp. 57, 59).

Oldham-Cooper, R. E., Hardman, C. A., Nicoll, C. E., Rogers, P. J., & Brunstrom, J. M. (2011). Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake. *The American Journal of Clinical Nutrition*, *93*(2), 308–313. https://doi.org/10.3945/ajcn.110.004580 (see p. 57).

Polivy, J., Herman, C. P., & Warsh, S. (1978). Internal and external components of emotionality in restrained and unrestrained eaters. *Journal of Abnormal Psychology*, *87*(5), 497–504. https://doi.org/10.1037/0021-843x.87.5.497 (see p. 64).

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/. (see p. 61).

Reichenberger, J., Kuppens, P., Liedlgruber, M., Wilhelm, F. H., Tiefengrabner, M., Ginzinger, S., & Blechert, J. (2018). No haste, more taste: An EMA study of the effects of stress, negative and positive emotions on eating behavior. *Biological Psychology*, *131*, 54–62 (see pp. 57, 58).

Robinson, E., Aveyard, P., Daley, A., Jolly, K., Lewis, A., Lycett, D., & Higgs, S. (2013). Eating attentively: A systematic review and meta-analysis of the effect of food intake memory and awareness on eating. *The American Journal of Clinical Nutrition*, *97*(4), 728–742. https://doi.org/10.3945/ajcn.112.045245 (see pp. 57, 65).

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02 (see p. 64).

Spence, C., & Shankar, M. U. (2010). The influence of auditory cues on the perception of, and responses to, food and drink. *Journal of Sensory Studies*, *25*(3), 406–430. https://doi.org/10.1111/j.1745-459x.2009.00267.x (see p. 57).

Stafford, L. D., & Dodd, H. (2013). Music increases alcohol consumption rate in young females. *Experimental and Clinical Psychopharmacology*, *21*(5), 408–415. https://doi.org/10.1037/a0034020 (see p. 57).

Stroebele, N., & de Castro, J. M. (2006). Listening to music while eating is related to increases in people's food intake and meal duration. *Appetite*, *47*(3), 285–289. https://doi.org/10.1016/j.appet.2006.04.001 (see pp. 57, 58).

Stutts, J., Feaganes, J., Reinfurt, D., Rodgman, E., Hamlett, C., Gish, K., & Staplin, L. (2005). Driver's exposure to distractions in their natural driving environment. *Accident Analysis & Prevention*, *37*(6), 1093–1101. https://doi.org/10.1016/j.aap.2005.06.007 (see p. 58).

Tang, D. W., Fellows, L. K., & Dagher, A. (2014). Behavioral and neural valuation of foods is driven by implicit knowledge of caloric content. *Psychological Science*, *25*(12), 2168–2176 (see p. 59).

Tobin, J. (1958). Liquidity preference as behavior towards risk. *The Review of Economic Studies*, *25*(2), 65. https://doi.org/10.2307/2296205 (see p. 62).

Torres, S. J., & Nowson, C. A. (2007). Relationship between stress, eating behavior, and obesity. *Nutrition*, *23*(11-12), 887–894. https://doi.org/10.1016/j.nut.2007.08.008 (see pp. 57, 58, 65).

van der Wal, R. C., & van Dillen, L. F. (2013). Leaving a flat taste in your mouth. *Psychological Science*, *24*(7), 1277–1284. https://doi.org/10.1177/0956797612471953 (see pp. 58, 65).

Voedingscentrum. (n.d.). Wat is de voedingswaarde van chips op basis van aardappelmeel. https://www.voedingscentrum.nl/nl/service/vraag-en-antwoord/gezonde-voeding-en-voedingsstoffen/hoeveel-voedingswaarden-zitten-erin/chips-op-basis-van-aardappelmeel.aspx (see p. 60).

Wallis, D. J., & Hetherington, M. M. (2004). Stress and eating: The effects of ego-threat and cognitive demand on food intake in restrained and emotional eaters. *Appetite*, *43*(1), 39–46. https://doi.org/10.1016/j.appet.2004.02.001 (see pp. 58, 65).

Wallis, D. J., & Hetherington, M. M. (2009). Emotions and eating. Self-reported and experimentally induced changes in food intake under stress. *Appetite*, *52*(2), 355–362 (see pp. 58, 65).

Yamauchi, K. (2013). *Gran turismo [video game]*. Sony Interactive Entertainment. (See p. 60).

Young, M. S., Mahfoud, J. M., Walker, G. H., Jenkins, D. P., & Stanton, N. A. (2008). Crash dieting: The effects of eating and drinking on driving performance. *Accident Analysis & Prevention*, *40*(1), 142–148. https://doi.org/10.1016/j.aap.2007.04.012 (see p. 58).

Zontone, P., Affanni, A., Bernardini, R., Del Linz, L., Piras, A., & Rinaldo, R. (2020). Stress evaluation in simulated autonomous and manual driving through the analysis of skin potential response and electrocardiogram signals. *Sensors*, *20*(9), 2494. https:/ / doi.org/ 10.3390/ s20092494 (see p. 58).

Supplementary Tables and Figures

Table 4 Means and standard deviations or *N* of gender, age and nationality as a function of condition (driving; control).

Condition*	Gender	Age	Nationality
Driving	15 M, 43 F	22.00 (2.66)	29 Dutch, 8 German, 3 English, 18 other
Control	15 M, 43 F	21.60 (3.02)	39 Dutch, 8 German, 11 other

**There were no statistically significant differences between the groups.*

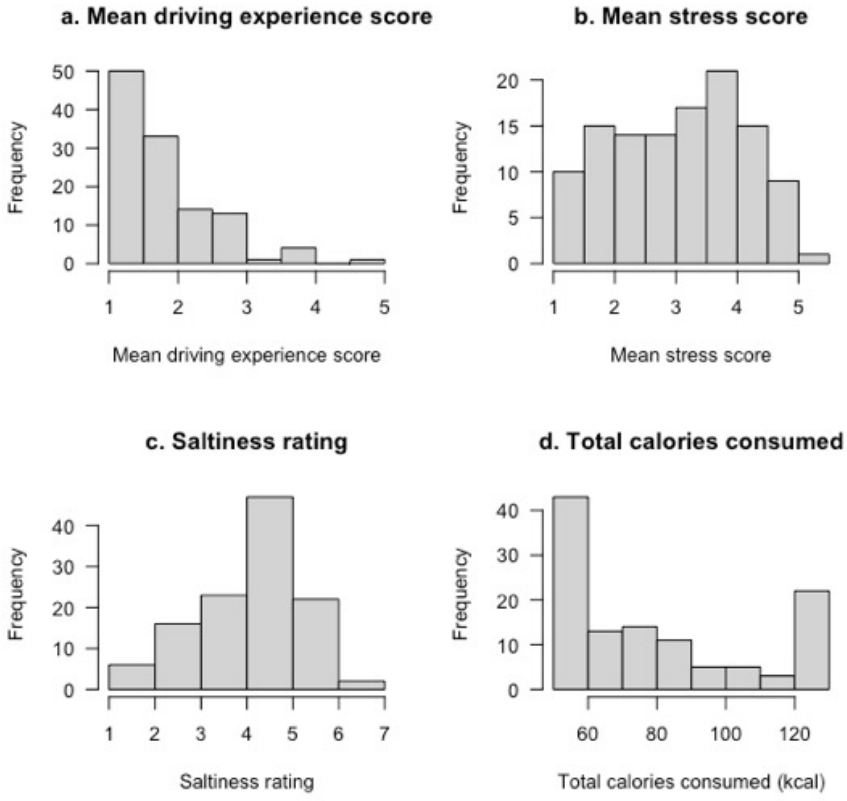


Figure 3 Histograms showing the distributions for a. mean Driving Experience, b. mean stress score, c. salt intensity rating, d. total amount of calories consumed.

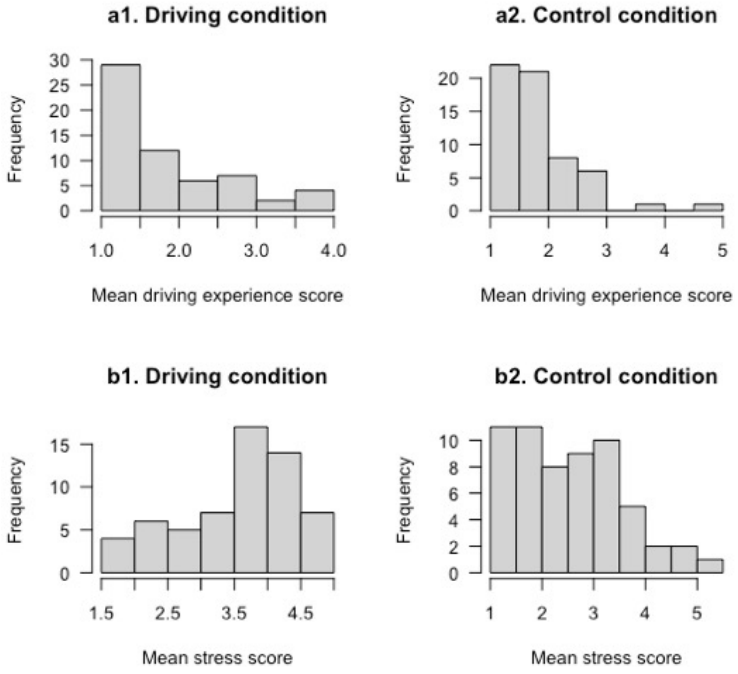


Figure 4 Histograms showing the distributions per condition (Driving and Control) for a. mean Driving Experience, b. mean stress score.

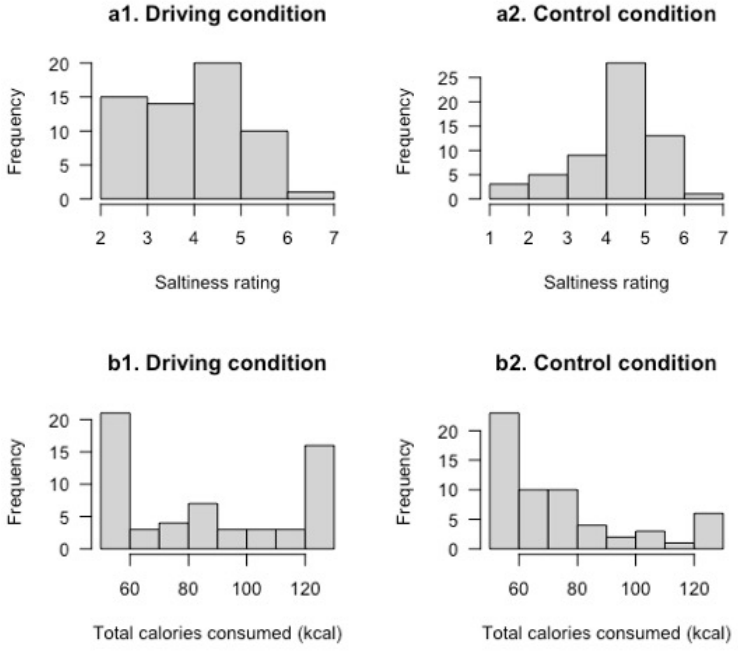


Figure 5 Histograms showing the distributions per condition (Driving and Control) for a. salt intensity rating, b. total amount of calories consumed.

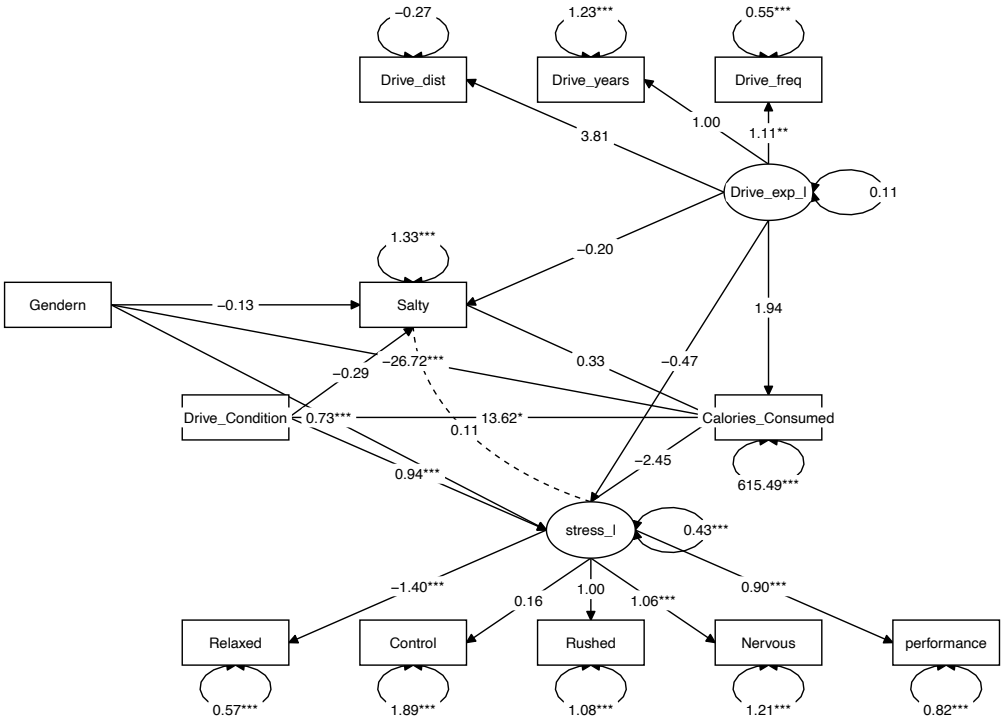


Figure 6 Structural Equation Modeling path model. Drive_dist = average kms driven in last year; Drive_freq = weekly driving frequency; Drive_years = years of having drivers' license; Drive_exp_l = latent variable of driving experience; Gendern = gender, male (0) or female (1); Salty = how salty the chips were perceived during the experiment; Drive_Condition = experimental condition, either completing a driving simulation (1) or control condition (0); Calories_consumed = the amount of calories consumed after the driving manipulation; stress_l = latent variable for stress experienced during the driving manipulation; Relaxed = how relaxed participants felt during the driving manipulation; Control = how in control participants felt during the driving manipulation; Rushed = how rushed participants felt during the driving manipulation; Nervous = how nervous participants felt during the driving manipulation; Performance = how well participants felt they performed during the driving manipulation.

Challenges of Using Signaling Data From Telecom Network in Non-Urban Areas

Håvard Boutera Toft[1,2], Alexey Sirotkin[3], Markus Landrø[1,2], Rune Verpe Engeset[1,2], Jordy Hendrikx[4,2]

Outdoor recreation continues to increase in popularity. In Norway, several avalanche fatalities are recorded every year, but the accurate calculation of a fatal accident rate is impossible without knowing how many people are exposed. We attempted to employ signaling data from telecom networks to enumerate backcountry travelers in avalanche terrain. Each signaling data event contains information about which coverage area the phone is connected to and a timestamp. There is no triangulation, making it impossible to know whether the associated phone is moving or stationary within the coverage area. Hence, it is easier to track the phone's movement through different coverage areas. We utilize this by enumerating the number of people with phones traveling to avalanche-prone terrain for the 2019-2020 winter season. We estimated that 13,666 phones were in avalanche terrain during the season, ranging from 0 to 118 phones per day with an average of 75 phones per day. We correlated the number of phones per day against amount of daylight (R^2=0.186, $p < 0.01$), weekends and holidays (R^2=0.073, $p < 0.01$), and number of bulletin views (R^2=0.045, $p < 0.01$). Unfortunately, the validation revealed discrepancies between the estimated positions in the mobile network and the true reference positions as collected with a GPS. We attribute this to the algorithm being designed to measure urban mobility and the long distance between the base transceiver stations in mountainous areas. This lack of coherence between the signaling data and GPS records for rural areas in Norway has implication for the utility of signaling data outside of urban regions.

Keywords *avalanche, risk, signaling data, telecom, non-urban areas*

[1]Norwegian Water Resources and Energy Directorate
[2]UiT The Arctic University of Norway
[3]Telia Company
[4]Antarctica New Zealand

Received
July 8, 2022
Accepted
December 14, 2022
Published
May 3, 2023
Issued
December 18, 2023

Correspondence
Norwegian Water Resources and Energy Directorate
htla@nve.no

The number of avalanche fatalities is generally well-documented (Thapa, **2010**; Willibald et al., **2019**), but obtaining a reliable measure of the total population (denominator) of people accessing avalanche terrain is difficult due to the open-access nature of these activities (Winkler et al., **2016**). However, there are multiple indirect proxies suggesting that backcountry travelers in avalanche terrain have increased in recent years (Birkeland et al., **2017**; Jekich et al., **2016**; Techel et al., **2016**; Winkler, **2015**). Backcountry travelers voluntarily expose themselves to avalanche risk during recreational activities such as skiing, snowboarding, snowshoeing, and snowmobiling (Johnson et al., **2020**).

If the entire population of backcountry travelers accessing avalanche terrain was known, it would be possible to calculate the likelihood of being killed by doing that activity in terms of micromorts. A micromort is a unit of risk, which denotes a one-in-a-million chance of death (Howard, **1984**). The calculation of micromorts is important as it would permit comparison to other recreational activities (e.g., skydiving, scuba diving and mountain biking) and a commensurate level of interventions, through targeted education and hazard awareness over time.

Several studies have tried to estimate the risk of death from recreational skiing, using such methods as rough estimates (Valla, **1984**), light barriers and counting at specific locations (Zweifel et al., **2006**), surveys (Sole & Emery, **2008**; Winkler et al., **2016**), and archived logs from mechanized skiing (Walcher et al., **2019**).

Take-home Message

We attempted to utilize signaling data to enumerate backcountry travelers in avalanche terrain. A representative sample would enable us to calculate the fatal accident rate. Unfortunately, the spatial validation revealed discrepancies between the estimated positions in the mobile network and the true reference positions collected with a GPS.

However, many of these methods only represent a crude measure of backcountry users for a small defined area, short time frame, or generalized survey data.

In Norway, an average of 6.5 avalanche fatalities have occurred per year over the last 10 years, but this has varied from 2 in the 2016-2017 winter season to 13 in the 2018-2019 season (Figure 1). While these fatalities provide some insight into avalanche risk, we are unable to estimate the fatality rate, as we do not have an estimate of the total number of people that expose themselves to this risk. Therefore, we are unable to assess if these changes in avalanche fatalities are due to changes in the number of people exposed, the snow cover, or the risk management. The latter is of great interest for avalanche forecasting services and educational institutions worldwide. Currently, no suitable methods exist to measure the effects of structured interventions, such as avalanche education or avalanche forecasting.

Furthermore, in the last 5 years, the trendline for avalanche fatalities has flattened out at approximately 6 fatalities (Figure 1, 10-year moving average). However, over this same period, we find it likely that there have been many more people in the mountains due to the increased popularity of backcountry travel. This increase is supported by various proxies, including the number of unique users accessing online avalanche forecasts (Engeset et al., **2018**). Therefore, does this increase in use and relatively steady count of fatalities suggest that the fatal accident rate has decreased over time? This is difficult to ascertain when we do not have a reliable base rate estimate of how many people are exposed to avalanche terrain every day or from year to year.

While our focus is on backcountry travelers in avalanche terrain, the same issue is shared by many other outdoor recreation activities, including but not limited to hiking, mountain biking, paragliding, trail running, and white-water kayaking. The fatalities and respective hazard-causing deaths are documented in all of these cases. The number of hours backcountry travelers expose themselves to avalanches, also known as the base rate, is absent (Johnson et al., **2020**; Kahneman & Tversky, **1973**). As such, a method to efficiently collect data on avalanche exposure is of value to the broader community of outdoor recreation.

Avalanches cause significant human and material losses (Schweizer, **2008**). Mitigation policies and prioritization require a qualitative basis from which to design strategies and allocate resources. WMO (**2021**) recommends a risk-based approach to warnings and mitigation (adopted by government agencies such as the Norwegian Water Resources and Energy Directorate) that requires base rate data. Due to the lack of exposure data, base rates are challenging to calculate in terms of people traveling in avalanche terrain. With base rate data, it is easier to understand which natural hazards need the most attention, the amounts of resources that are needed, and which measures are most efficient from a cost-benefit perspective.

The base rate information could also be used to validate whether an increase in objective danger correlates with avalanche danger levels. Winkler et al. (**2021**) calculated a relative risk between the danger levels, but without a base rate, they could not calculate the absolute risk (i.e., micromorts). Furthermore, without a valid base rate measure, Bayesian approaches, which utilize diagnostic tests (also known as stability tests) to assess avalanche decision-making, lack important input data (Ebert, **2019**; Techel et al., **2020**).

Given the ubiquitous use of mobile phones in Norway (Statista, **2021**), with 99.9% of the population having access to 4G coverage (MLGM, **2021**), there is a potential opportunity to obtain some insight into the total exposure to avalanche terrain. Telia, one of the largest mobile network operators (MNOs) in Norway, collects a vast amount of anonymized data through what is referred to as signaling data. Every time a phone communicates with a base transceiver station (BTS) (e.g., a phone

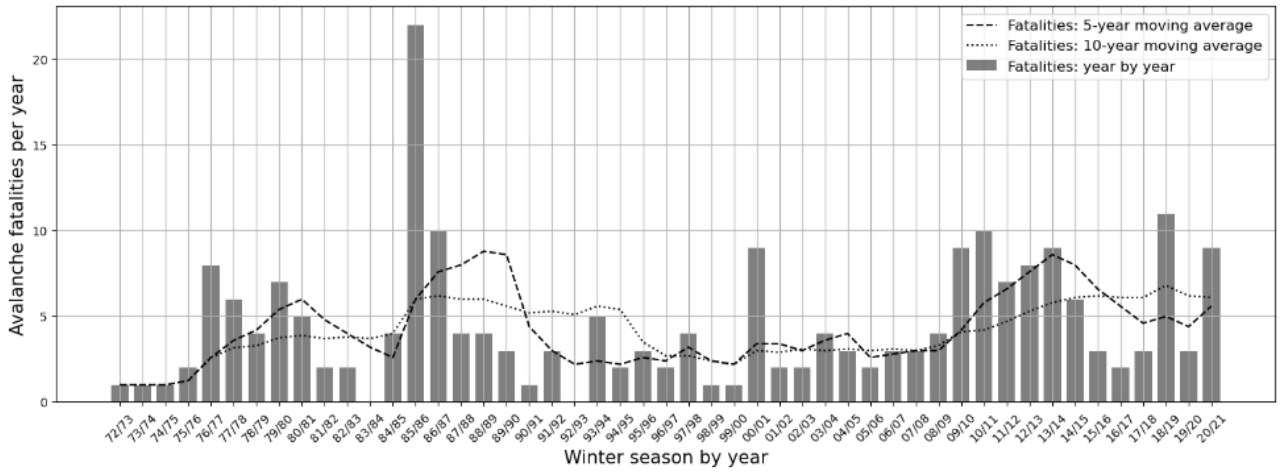


Figure 1 Recreational avalanche fatalities in Norway by winter season from 1972 to 2021 with a 5 and 10-year moving average (NGI, **2019**; Varsom, **2021**).

call, text message, or the phone itself checks for new emails), signaling data is generated. On average, a Telia subscriber generates around 300-400 active and passive signaling events a day, or roughly 15 events per hour. The vast amount of data collected makes it an appealing data source when studying human mobility (Zhao et al., **2016**).

During the last few decades, telecom data have been widely used in the research community. Many useful findings of human activity have been reported for urban areas (González et al., **2008**; Song et al., **2010**). To our knowledge, there is no research applying telecom data in non-urban areas other than Francisco et al. (**2018**). The reason for this could be the relatively lower density of BTSs in rural and mountainous terrain, with the majority located where people live, work, and travel (Zhao et al., **2016**).

Norway has a vast number of remote mountains, fjords, and islands. It is also among the least densely populated countries globally, with a population density of 15 people per square kilometer (UN, **2021**). Despite this, the MNOs in Norway have been ranked among the top 10 providers worldwide with respect to cell phone coverage for several years in a row (Speedtest, **2021**). As a result of the excellent coverage, most mountainous areas in Norway have full 4G coverage (Telenor, **2021**; Telia, **2021**), and therefore their signaling data are expected to have some utility in these areas.

In this study, we attempted to use anonymized and aggregated signaling data to count how many people expose themselves to avalanche terrain around Tromsø, Norway. We selected this area as historically, nearly 2/3 of all recreational avalanche fatalities in Norway occur in this county (Varsom, **2021**). However, because no one has been able to accurately estimate how many people enter avalanche terrain in this region, it is impossible to say whether this high number of fatalities is solely due to more users in the area, or if it is more dangerous to ski in the area around Tromsø compared to the rest of the country. Without the base rate information, we are unable to determine which of the two hypotheses is correct (Johnson et al., **2020**; Kahneman & Tversky, **1973**).

Secondly, we also want to use this method to help assess whether the fatal accident rate (FAR) from avalanches has increased or decreased during the last decade. Despite the number of avalanche fatalities over the last ten years having been relatively stable, there is a general agreement that there has been a significant increase in traffic amongst different groups of backcountry travelers in the same period. This view of increasing use is supported by a range of proxies, including the

number of people seen in avalanche terrain, the number of vehicles at trailheads, and the sale of backcountry traveling equipment. This increase of use, combined with a relatively stable fatality count, suggests that the FAR has decreased during the last few decades (Techel et al., **2016**).

The challenges of determining the number of people exposing themselves to avalanche risk in the backcountry and calculating the risk of skiing in avalanche terrain, have been approached by several others using a range of imperfect methods. For example, Zweifel et al. (**2006**) used light barriers and voluntary registration boards at four sites near Davos, Switzerland. Using these methods, Zweifel et al. (**2006**) calculated the individual risk factor for this population and found it lower than the risk of driving a car. However, this was for very limited area of Switzerland, and represents an engaged and self-selecting audience that voluntarily provide registration information. There have also been several studies using GPS-tracking and surveys to assess terrain use (Buhler & Floyer, **2016**; Hendrikx & Johnson, **2014**; Hendrikx et al., **2016**; Sykes et al., **2020**; Thumlert & Haegeli, **2017**; Winkler et al., **2021**), but these studies are not representative for the whole population and are generally skewed towards more engaged and advanced users. Passive tracking of backcountry users with time-lapse camera technology has also been used (Saly et al., **2020**), but was also limited to a small geographic area. The use of telecom data for avalanche terrain is limited to a single study by Francisco et al. (**2018**), who undertook a case study to track backcountry users in the Sorteny valley, Andorra. They obtained access to raw call detail records (CDRs), including an estimated position for each record with an accuracy of 150 meters for a period of 20 days. From these CDRs, they created daily frequency plots and compared them with avalanche danger, temperature, wind, snow depth, solar radiation, and precipitation. Unfortunately, Francisco et al. (**2018**) did not provide any information regarding how the position (latitude, longitude) was established or validated.

Our study attempted to build on these prior studies and used truly anonymized signaling data from Telia Company to count the total number of backcountry users within one avalanche forecasting zone in northern Norway. We also explored how these counts changed in relation to known drivers of usage, including weekends and holidays, and variable environmental conditions.

Methods

Telia uses telecom network data, which is one of the most extensive and continuously generated datasets in society today. The network data exceeds billions of data points every day in each Nordic country. These are stored in the Telia database for billing, network optimization, and other purposes. However, in contrast to regular data services, Telia can safeguard that no individuals can be identified in the dataset, while still providing extrapolated national movement patterns that are statistically representative for the entire population and not just Telia subscribers.

Using signaling data, Telia can produce mobility insights through a GDPR-compliant method. They do this by never storing, processing, or exposing data that can identify an individual, and the smallest result generated is in groups of 5 individuals within the same movement chain (Ågren et al., **2021**).

Telia does not have the exact position of each phone in their signaling data, and new data is only generated when the phone is actively or passively used (i.e., calling, SMS, transfer of data), but most smartphones today are constantly checking for updates, and thus constantly generating signaling data.

Each signaling data record includes a timestamp and the coverage area (Cell ID) to which the phone is connected. The best server estimate (BSE), which is the estimated coverage area, is defined for each Cell ID. Most BTSs have several Cell IDs due to the different antennas pointing in diverging directions. Thus, the Cell ID provides more specific information about the position of the phone than only using the BTS. The BSE consists of uniquely shaped polygons representing the coverage area of each Cell ID. The MNOs collect a lot of data, but the utility of that data for research purposes is limited due to privacy concerns. Telia Company does not use any triangulation methodology to define a more exact location due to their strict privacy policy, but by analyzing the data over time, it is possible to generate movement chains from the signaling data. Telia's

algorithms process the movement chains to form insights. They were originally intended for urban areas, but we have employed them to assess whether we can count skiers' phones in avalanche terrain using the insights from signaling data.

The algorithms that process the movement chains are designed to capture three different patterns. The overview below intends to provide a working understanding of how Telia distill relevant data for each report.

1. Activity report – where crowds are spending time without directional movement.
2. Routing report – where crowds are passing by without stopping.
3. Origin-destination report – the trips made by crowds between origins and destinations.

In this study, we utilize the activity report, which captures how many subscribers spend a defined amount of time in a defined geographical area. The activity report can be produced from a regional level and down to a statistical grid, with the lowest resolutions being 500 x 500 meters in a dense urban area. The resolution is flexible, so the grids are larger to secure GDPR compliance in rural settings. It is possible to filter the amount of time spent in a defined area, or use timestamps to reveal when visitors arrived or left an area during the day.

The spatial resolution of mobile network data is dependent on the size of the Cell ID that the cellphone has been communicating with. Each BTS has several Cell IDs with a geographic coverage area. When a device moves around it will connect to multiple different Cell IDs, leaving a movement chain. The initial processing involves turning this raw data trace into dwells (activities taking place in one location) and movements (Figure 2).

To utilize this methodology, we defined a geographical area for the avalanche terrain. We also defined where people live (i.e., populated areas) to identify areas that we could distinguish between avalanche terrain and populated areas. We defined populated areas and avalanche terrain on a map using GIS software (Figure 3). Definitions and methods for defining these areas are outlined in the sections below.

Populated areas

Statistics Norway (**2021**) has created a GIS layer with the number of inhabitants per 1x1 square kilometers. We used this layer and defined populated areas as grid cells with more than ten inhabitants.

Avalanche terrain

Avalanche terrain can be defined using the Avalanche Terrain Exposure Scale (ATES) framework (Statham et al., **2006**). Using the nationwide ATES layer developed by Larsen et al.(2020), we defined avalanche terrain as the sum of simple, challenging, and complex avalanche terrain. Numerous houses and roads lie within avalanche terrain (Kalsnes et al., **2021**). We removed all avalanche terrain within 300 meters of a house or a road from the GIS layer. The distance of 300 meters was chosen to avoid counting people that are driving a car or living in a house, but not moving between a populated area and avalanche terrain.

Mobility analysis

The two layers with populated areas and avalanche terrain were exported and shared with Telia. They applied the layers with their BSE of the coverage area and identified areas where it was possible to distinguish between populated regions (purple) and potential avalanche terrain (red) (Figure 4). Using the insights from the movement chains, Telia counted how many phones traveled into avalanche terrain using signaling data.

Given the nature of the terrain, the most common backcountry trips around Tromsø include a vertical elevation gain of between 800-1200 vertical meters. Assuming a regular uphill pace of 400-600 vertical meters per hour, this could cause uphill travel to take as few as 2 hours for the fittest recreationists. Most people also hike and ride during the daytime. Therefore, we added a filter that only kept phones that were in avalanche-prone terrain for at least 2 hours between 07.00-23.00, during the 2019-2020 avalanche forecasting season (1st of December until 31st of May). This period includes the spring season when the Covid-19 pandemic started. Large parts of Norway closed down on March 13th and there is likely a drop in tourists after this date.

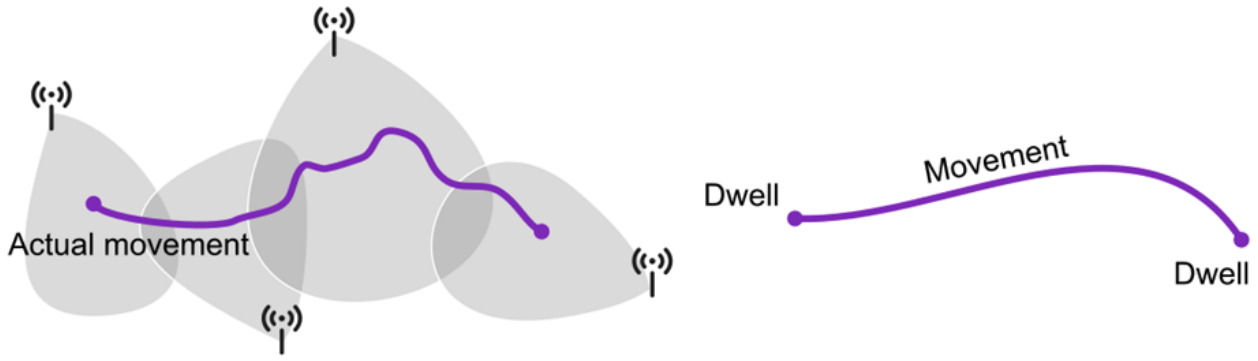


Figure 2 The movement of each cellphone could be tracked through different coverage areas.

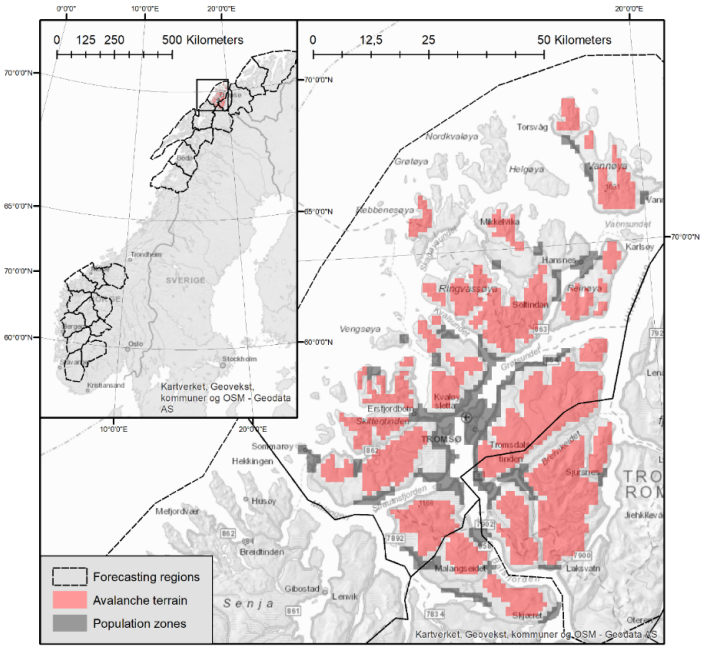


Figure 3 The case study area Tromsø, Northern Norway. Avalanche terrain is colored in red, while the populated areas are colored in dark gray. The avalanche forecast regions are delineated using a dashed line on the inset map.

Validation

To improve the insights from the movement chains, Telia has developed an algorithm that can assign the most likely position within the Cell ID. Telia has targeted the algorithm against normal behavior, which means that the positions will be biased towards populated areas and roads where most people travel. The in-depth details regarding the algorithm are considered a trade secret and are not disclosed due to Telia's commercial interests. Using the output from the algorithm enables us to compare the GPS position to signal data-derived position. The GPS on their watch has a position accuracy of 5-10 m (Wing et al., **2005**).

Correlation with other usage factors

We correlated the number of people per day against the amount of daylight, number of avalanche bulletin page views, weekends and holidays, the daily avalanche danger, percentage of cloud cover, wind strength, and precipitation. All weather parameters were aggregated between 07.00 in the morning and 23.00 in the evening to only account for the conditions during daytime. Amount of daylight was calculated for a latitude of 69° North (SatAgro, **2019**). The daily avalanche danger level was provided by Varsom (**2021**) and the number of page views for the avalanche forecast on Varsom.no was provided by Google Analytics.

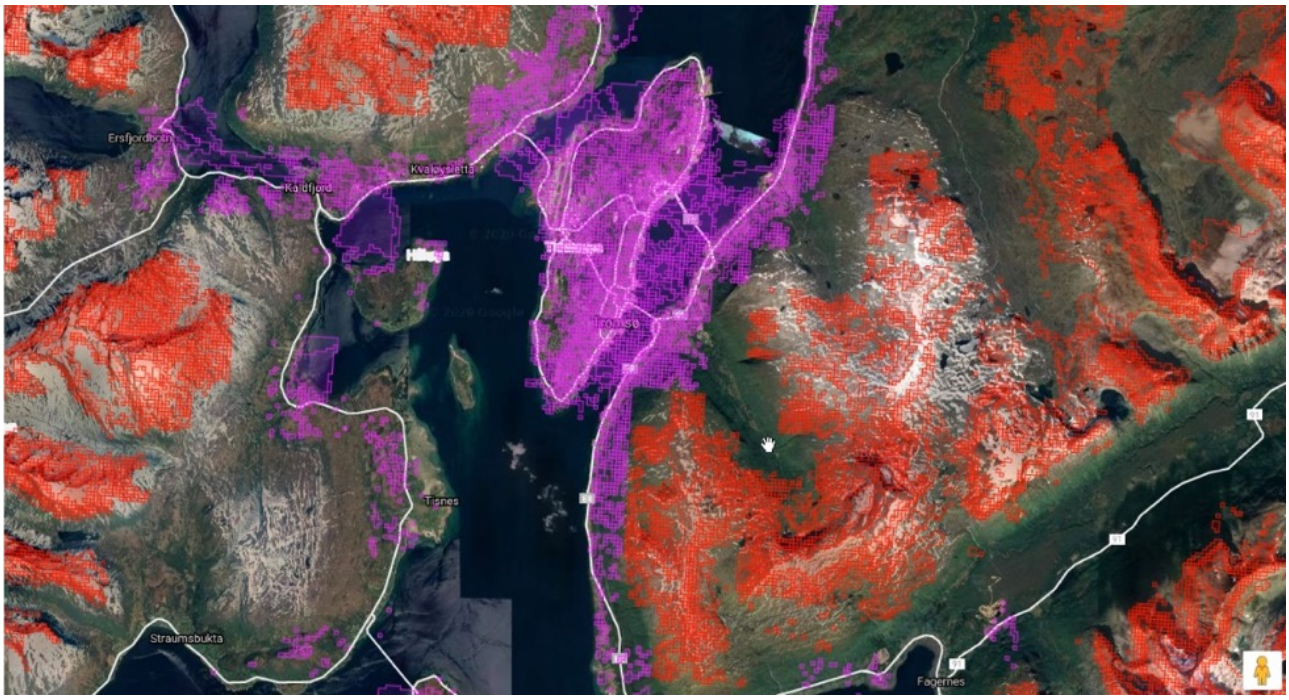


Figure 4 Example of identified areas around Tromsø where the Telia could distinguish between populated areas (purple) and potential avalanche terrain (red) given their BTS coverage in the region.

Table 1 Number of people in potential avalanche terrain versus different variables that could be controlling number of people in avalanche terrain. * Variable is not significant.

Number of people per day versus:	R^2	p-value
Amount of daylight	0.186	< 0.01
Weekend and holidays	0.073	< 0.01
Avalanche forecast page views	0.045	< 0.01
Avalanche bulletin	0.007	0.244*
Cloud cover	0.004	0.374*
Wind strength	0.002	0.521*
Precipitation	0.000	0.917*

Weekends and holidays were encoded as binary values of 0 or 1, with weekends and holidays coded as 1 and weekdays coded as 0. The weather variables were downloaded from the Norwegian Centre for Climate Services (**2021**) on an hourly basis.

Results

Mobility analysis

Using the mobility analysis methods, we estimated that 13,666 people were in avalanche terrain for at least two hours during the 2019-2020 season (December 1st, 2019, to May 31st, 2020, consisting of 182 days). The number of people in avalanche terrain per day varied from 0 to 118, with an average of 75 people per day.

Amount of daylight had the strongest, albeit very low, correlation (R^2 = 0.186, p < 0.01), followed by weekends and holidays (R^2 = 0.073, p < 0.01) and the number of forecast page views (R^2 = 0.045, p < 0.01). We also correlated against precipitation, wind, daily avalanche danger and cloud cover, but none of these parameters were statistically significant (Table 1).

Positional Validation

Using a phone with a special SIM card that was whitelisted (i.e., not anonymized in the

Table 2 Minimum, maximum, median, and 95% of all point distances between GPS track and signaling data spatial locations.

	Min	Max	Median	95% of points within	N (samples)
Trip 1	455 m	8,216 m	4,188 m	7,580 m	74
Trip 2	7,876 m	21,502 m	13,607 m	20,424 m	93
Trip 3	19 m	16,213 m	2,596 m	14,212 m	135
Trip 4	1,997 m	8,919 m	6,911 m	8,736 m	114
All trips	19 m	21,502 m	6,523 m	12,920 m	416

telecom network—users gave specific consent for this), our validation focused on the positional accuracy of the signaling data relative to the synchronous GPS records. When we compared these, we discovered that there was a discrepancy between the two data sets. In the examples (Figure 5), we can see that the estimated positions from the signaling data does not resemble the GPS track. Most of the signaling data positions are estimated to be in the valley bottom, following road corridors or out on the fjords. For all four trips, the positional difference ranged from 19 meters to 21,502 meters. The median positional difference was 6,523 meters and 95% of the points were within 12,920 meters (Table 2).

Discussion

A qualitative review of the four GPS tracks and the signaling data estimated locations shows discrepancies in the estimated positions from the two data sets as shown in Figure 4. This is further supported by our quantitative analysis, where all trips were off by several hundred meters to several kilometers (Table 2). Clearly these positional results are disappointing, and in strong contrast to the reported 150-meter accuracy of the geolocation in mountainous terrain in Andorra (Francisco et al., **2018**). It is difficult to directly compare our results regarding accuracy given that we do not know how Francisco et al. estimated their positional data, or how they validated the accuracy of the signaling data. The differences could be due to several factors, including the potential lower density of BTSs in Troms and/or the algorithm in Norway being designed by Telia for use in urban areas. By comparison, Jansen et al. (**2021**) found the position accuracy of telecom data to be roughly 500 meters in the cities and 3,000 meters in rural areas.

To validate our data, we wanted to check whether the Telia's algorithm estimated the correct locations in rural areas where the coverage areas for each Cell ID are much larger. The algorithm is tuned to work in populated areas where the coverage areas for each Cell ID are small, which makes it easier to estimate the position moving through different coverage areas. The difference in density of BTSs was one of the significant uncertainties in our study. After sending mountain runners out with whitelisted phones, we learned that the positioning of each phone did not work as well as we had initially hoped. When whitelisted phones were compared with actual GPS tracks, we found that the signaling data-derived locations would follow the road corridors leading to the mountains. When our mountain runners parked their cars at the foot of the mountain and started running up, the estimated position stayed in the valley bottom or out on the fjords. We quickly learned that what we initially believed to be a good dataset of ski traffic in the region from the signaling data was biased by the large coverage area of each Cell ID outside the cities. We think there are two primary reasons for this:

1. Telia's algorithm is targeted using data from people travelling on roads between houses, work, stores, etc.
2. The coverage area outside the populated areas is too large to define whether people are up in the mountains or not.

In the bigger picture, these problems are not that surprising. Mobile networks are built and optimized for urban areas where most people live, work and travel. Telecom companies specifically design and build their networks to cover large areas with the fewest possible number of antennas. We are trying to achieve the opposite, capturing signaling data from unpopulated areas where people usually do not travel due to lack of infrastructure. In a broader sense, this is the main limitation of our ability to accurately estimate the position of each phone in rural areas.

We also compared the data with parameters we expected would affect the number of people out in the mountains to initially verify our data. The parameters were amount of daylight (expected positive correlation), number of bulletin page views (expected positive correlation), the occurrence of weekends and holidays (expected positive correlation), rain (expected negative correlation), cloud cover (expected negative correlation), wind strength (expected negative correlation), and avalanche danger (expected negative correlation). For weather data, we only investigated data between 07.00-23.00 because this was the period, we counted people and would likely affect the decision to go skiing or not. The most important coefficient was amount of daylight,

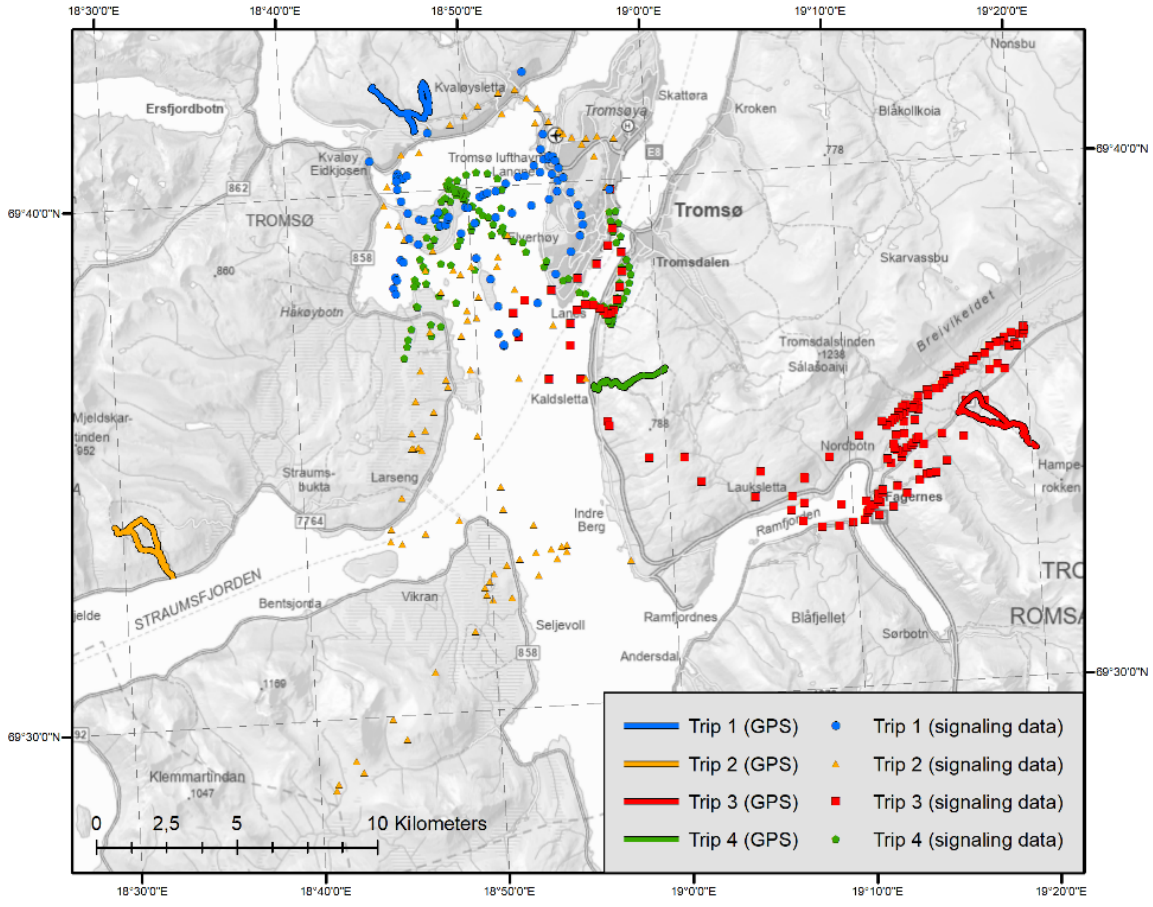


Figure 5 Comparison of 4 different color-coded trips using GPS data (line) and estimated positions from the signaling data (circle, triangle, square and pentagon).

followed by weekend/holidays and number of bulletin page views. The various weather parameters were not significant and had very low coefficient scores. The results are logical because most backcountry travelers are outside when there is daylight in the Arctic, but we had hoped for a better fit towards the weather parameters. This lack of fit in our simple correlations is most likely due to the inaccurate location positions from the signaling data, resulting in the additional counts of users that were not in avalanche terrain, but were included in our data set. This resulting data set is therefore much noisier and includes people in other areas outside of the immediate populated areas, but not necessarily in avalanche terrain.

Original Purpose

The objective of this manuscript is to document our attempts to use signaling data from telecom networks to count the number of backcountry travelers in avalanche terrain within the Tromsø avalanche forecasting region in Northern Norway. If this method had permitted an accurate count of people in avalanche terrain, we would have been able to obtain a very representative sample of overall terrain usage. Combined with the observed number of fatalities in the same avalanche forecasting region, we would also have been able to calculate how risky the activity is in terms of micromorts. Furthermore, we could have tracked terrain usage over several winter seasons to obtain data to assess whether there are any trends concerning the number of people accessing avalanche terrain over time and whether increased interventions, including the uptake of avalanche awareness courses and improved avalanche forecasting, is evident in a change in micromorts over time. To our knowledge, there have not been any studies utilizing telecom network data in non-urban areas. We believe that our results are significant for a broader audience to showcase potential pitfalls when using this type of data. We also highlight the importance of validating this type of data.

Limitations

As already noted in our discussion above, the positional accuracy of the signaling data when compared to the GPS data is the main limitation to the use of this methodology as currently presented. Access to the raw data, prior to analysis by the algorithm, which is targeted for urban use, might alleviate some of these issues, but this was considered outside the scope of the current study.

Furthermore, the reliability of any mobile phone tracking in avalanche terrain depends on users leaving their phone turned on for the duration of their trip. Many backcountry travelers elect to turn their cell phones off to purposefully save battery power for emergency calls. Travelers are also generally encouraged to turn their cell phones off or to flight mode to prevent potential interference with avalanche transceivers. This reality was reflected in a winter backcountry survey by Ortega et al. (**2018**) in Alaska, which showed that of the 63 users interviewed, approximately half of them typically leave their phone turned on whereas the rest turn theirs off or to flight mode.

The main limitation in making telecom data viable for counting people in avalanche-prone terrain is the lack of numerous BTSs in mountainous areas. A more specific algorithm could improve the data quality for this use case, but the BTS density is likely the key factor that would make the method more viable if a mountainous area with a higher density of BTSs is found.

Conclusion

In urban areas, each BTS with several Cell IDs is close together, which means that Telia can estimate more accurate positions given the small coverage area for each Cell ID. Even though Norway has exceptional cellphone coverage compared to many other countries, it is still insufficiently dense in our non-urban and mountainous study area case study. The long distances between the BTSs, and therefore large coverage areas, combined with the populated area-targeted algorithm, are the most likely reasons for the inability to accurately calculate the position of each phone in avalanche terrain. The poor correlation between the GPS track and the position of the whitelisted

phones means that we cannot trust the positional accuracy of this initial dataset as provided by Telia. Future work should focus on making a model that is independent of where most people travel. This study provides a useful, yet unsuccessful, case study that demonstrates the limits of signaling data for use in non-urban mountainous areas. It has relevant implications for the application of signaling data tracking to other outdoor recreation activities. We highlight the importance of validating positional data from signaling data to be used in mobility studies in remote areas.

Data

The data that support the findings of this study are available at https://zenodo.org/record/7891581.

Funding

This work is financially supported by the Norwegian Water Resources and Energy Directorate and Telia Company, who have generously provided dedicated working hours for the project.

Conflict of interest

The authors declare that they have no conflict of interest.

References

Ågren, K., Bjelkmar, P., & Allison, E. (2021). The use of anonymized and aggregated telecom mobility data by a public health agency during the COVID-19 pandemic: Learnings from both the operator and agency perspective. *Data & Policy*, *3*. https://doi:10.1017/dap.2021.11 (see p. 75).

Birkeland, K. W., Greene, E. M., & Logan, S. (2017). In response to avalanche fatalities in the united states by jekich et al. *Wilderness and Environmental Medicine*, *28*(4), 380–382. https://doi.org/10.1016/j.wem.2017.06.009 (see p. 72).

Buhler, R., & Floyer, J. (2016). *Using crowdsourced data to understand terrain usage patterns of backcountry recreational users*. International Snow Science Workshop. (See p. 75).

Ebert, P. A. (2019). Bayesian reasoning in avalanche terrain: A theoretical investigation. *Journal of Adventure Education and Outdoor Learning*, *19*(1), 84–95. https://doi.org/10.1080/14729679.2018.1508356 (see p. 73).

Engeset, R., Pfuhl, G., Landrø, M., Mannberg, A., & Hetland, A. (2018). Communicating public avalanche warnings – what works? *Natural Hazards and Earth System Science*, *18*, 2537–2559. https://doi.org/10.5194/nhess-18-2537-2018 (see p. 73).

Francisco, G., Apodaka, J., Travesset-Baro, O., Vilella, M., Margalef, A., & Pons, M. (2018). *Exploring the potential of mobile phone data (call detail records) to track and analyze backcountry skiers' dynamics in avalanche terrain*. International Snow Science Workshop. (See pp. 74, 75, 79).

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782. https://doi.org/10.1038/nature06958 (see p. 74).

Hendrikx, J., & Johnson, J. (2014). *Using global crowdsourced data to understand travel behavior in avalanche terrain*. International Snow Science Workshop. (See p. 75).

Hendrikx, J., Johnson, J., & Shelly, C. (2016). Using GPS tracking to explore terrain preferences of heli-ski guides. *Journal of Outdoor Recreation and Tourism*, *13*. https://doi.org/10.1016/j.jort.2015.11.004 (see p. 75).

Howard, R. A. (1984). On fates comparable to death. *Management Science*, *30*(4), 407–422 (see p. 72).

Jansen, R., Kovacs, K., Esko, S., Saluveer, E., Sõstra, K., Bengtsson, L., Li, T., Adewole, W. A., Nester, J., Arai, A., & Magpantay, E. (2021). Guiding principles to maintain public trust in the use of mobile operator data for policy purposes. *Data & Policy*, *3*. https://doi.org/10.1017/dap.2021.21 (see p. 79).

Jekich, B. M., Drake, B. D., Nacht, J. Y., Nichols, A., Ginde, A. A., & Davis, C. B. (2016). Avalanche fatalities in the united states: A change in demographics. *Wilderness & Environmental Medicine*, *27*(1), 46–52. https://doi.org/10.1016/j.wem.2015.11.004 (see p. 72).

Johnson, J., Mannberg, A., Hendrikx, J., Hetland, A., & Stephensen, M. (2020). Rethinking the heuristic traps paradigm in avalanche education: Past, present and future. *Cogent Social Sciences*, *6*(1), 1807111. https://doi.org/10.1080/23311886.2020.1807111 (see pp. 72, 73, 74).

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237–251. https:/ / doi.org/ 10.1037/ h0034747 (see pp. 73, 74).

Kalsnes, B., Solheim, A., Sverdrup-Thygeson, K., Dingsør-Dehlin, F., Wasrud, J., Indrevær, K., & Bergbjørn, K. (2021) (see p. 76).

MLGM. (2021). Bredbånd og mobil. In *Ministry of local government and modernization*. https://www

.regjeringen.no/no/tema/transport-og-kommunikasjon/elektronisk-kommunikasjon/ekomartikler_2019/bredband-og-mobil/id2642610/ (see p. 73).

NGI. (2019). *Ulykker med død*. Norwegian Geotechnical Institute. https://www.ngi.no/Tjenester/Fagekspertise/Snoeskred/snoskred.no2/Ulykker-med-doed (see p. 74).

Norwegian Centre for Climate Services. (2021). *Observasjoner og værstatistikk*. https://seklima.met.no/ (see p. 78).

Ortega, C., Wollgast, R., & Latosuo, E. (2018). Presence of social media use and smart phone technology among backcountry skiers and snowboarders, hatcher pass, alaska. *Proceedings of the International Snow Science Workshop*, 1583 (see p. 81).

Saly, D., Hendrikx, J., Birkeland, K. W., Challender, S., & Johnson, J. (2020). Using time lapse photography to document terrain preferences of backcountry skiers. *Cold Regions Science and Technology*, *172*, 102994. https://doi.org/10.1016/j.coldregions.2020.102994 (see p. 75).

SatAgro. (2019). Suntime. *GitHub*. https://github.com/SatAgro/suntime (see p. 77).

Schweizer, J. (2008). Snow avalanche formation and dynamics. *Cold Regions Science and Technology*, *54*(3), 153–154. https://doi.org/10.1016/j.coldregions.2008.08.005 (see p. 73).

Sole, A., & Emery, C. (2008). Human risk factors in avalanche incidents (see p. 72).

Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics*, *6*(10), 818–823. https://doi.org/10.1038/nphys1760 (see p. 74).

Speedtest. (2021). Best mobile coverage 2021. *Speedtest Awards*. https://www.speedtest.net/awards/coverage/? (see p. 74).

Statham, G., McMahon, B., & Tomm, I. (2006). The avalanche terrain exposure scale. *International Snow Science Workshop Proceedings, Telluride*, 491–497. http://www.lauegi.conselharan.org/files/ATES (see p. 76).

Statista. (2021). Share of individuals who had access to a smartphone in norway from 2012 to. https://www.statista.com/statistics/631747/norway-smartphone-user-penetration/ (see p. 73).

Statistics Norway. (2021). *Kart og geodata fra SSB*. https://www.ssb.no/en (see p. 76).

Sykes, J., Hendrikx, J., Johnson, J., & Birkeland, K. W. (2020). Combining GPS tracking and survey data to better understand travel behavior of out-of-bounds skiers. *Applied Geography*, *122*. https://doi.org/10.1016/j.apgeog.2020.102261 (see p. 75).

Techel, F., Jarry, F., Kronthaler, G., Mitterer, S., Nairz, P., Pavšek, M., Valt, M., & Darms, G. (2016). Avalanche fatalities in the european alps: Long-term trends and statistics. *Geographica Helvetica*, *71*(2), 147–159. https://doi.org/10.5194/gh-71-147-2016 (see pp. 72, 75).

Techel, F., Winkler, K., Walcher, M., van Herwijnen, A., & Schweizer, J. (2020). On snow stability interpretation of extended column test results. *Natural Hazards and Earth System Sciences*, *20*(7), 1941–1953. https://doi.org/10.5194/nhess-20-1941-2020 (see p. 73).

Telenor. (2021). Dekningskart. *Telenor*. https://www.telenor.no/dekning/#dekningskart (see p. 74).

Telia. (2021). Dekningskart. *Telia*. https://www.telia.no/nett/dekning/ (see p. 74).

Thapa, B. (2010). The mediation effect of outdoor recreation participation on environmental attitude-behavior correspondence. *The Journal of Environmental Education*, *41*(3), 133–150. https://doi.org/10.1080/00958960903439989 (see p. 72).

Thumlert, S., & Haegeli, P. (2017). Describing the severity of avalanche terrain numerically using the observed terrain selection practices of professional guides. *Natural Hazards*, 1–27. https://doi.org/10.1007/s11069-017-3113-y (see p. 75).

UN. (2021). *Country profile*. Norway. United Nations. https://data.un.org/CountryProfile.aspx/_Images/CountryProfile.aspx?crName=Norway (see p. 74).

Valla, F. (1984). The french experience in avalanche education for skiers. *International Snow Science Workshop Proceedings*, 70–77 (see p. 72).

Varsom. (2021). Norwegian avalanche fatalities. https://www.varsom.no/ulykker/snoskredulykker-og-hendelser/ (see pp. 74, 77).

Walcher, M., Haegeli, P., & Fuchs, S. (2019). Risk of death and major injury from natural winter hazards in helicopter and snowcat skiing in canada. *Wilderness & Environmental Medicine*, *30*(3), 251–259. https://doi.org/10.1016/j.wem.2019.04.007 (see p. 72).

Willibald, F., van Strien, M. J., Blanco, V., & Grêt-Regamey, A. (2019). Predicting outdoor recreation demand on a national scale – the case of switzerland. *Applied Geography*, *113*, 102111. https://doi.org/10.1016/j.apgeog.2019.102111 (see p. 72).

Wing, M., Eklund, A., & Kellogg, L. D. (2005). Consumer-grade global positioning system (GPS)

accuracy and reliability. *Journal of Forestry*, *103*(4), 169–173 (see p. 77).

Winkler, K. (2015). Entwicklung des lawinenrisikos bei aktivitäten im freien gelände. In *Lawinen und recht. tagungsband zum internationalen seminar* (pp. 109–112). (See p. 72).

Winkler, K., Fischer, A., & Techel, F. (2016). Avalanche risk in winter backcountry touring: Status and recent trends in switzerland. *Zurich Open Repository and Archive*, 270–276. https://doi.org/10.5167/uzh-126510 (see p. 72).

Winkler, K., Schmudlach, G., Degraeuwe, B., & Techel, F. (2021). On the correlation between the forecast avalanche danger and avalanche risk taken by backcountry skiers in switzerland. *Cold Regions Science and Technology*, *188*, 103299. https://doi.org/10.1016/j.coldregions.2021.103299 (see pp. 73, 75).

WMO. (2021). WMO guidelines on multi-hazard impact-based forecast and warning services: Part II: Putting multi-hazard IBFWS into practice. *WMO*, (1150) (see p. 73).

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738–1762. https://doi.org/10.1080/13658816.2015.1137298 (see p. 74).

Zweifel, B., Raez, A., & Stucki, T. (2006). *Avalanche risk for recreationists in backcountry and in off-piste area: Surveying methods and pilot study at davos*. International Snow Science Proceedings. (See pp. 72, 75).

Empathic Accuracy, Mindfulness, and Facial Emotion Recognition: An Experimental Study

Marije aan het Rot[1,2], Merle-Marie Pittelkow[1], D. Elisabeth Eckardt[1], Nils Simonsen[1], Brian D. Ostafin[1]

Background and Objectives: Empathic accuracy, i.e., the degree to which one is able to accurately infer the emotions of others, may be acutely malleable. We examined this idea by testing the immediate effect of a brief mindfulness intervention or facial emotion recognition training. **Methods:** Participants were English- or Dutch-speaking psychology students who were assigned to one of three brief intervention conditions (all instructions given in English): (1) verbal instructions for practicing awareness of their body (mindfulness, n = 23); (2) verbal and visual instructions regarding the detection of visual cues for anger, fear, sadness, and happiness (facial emotion recognition training, n = 23); or (3) a verbal, neutral didactic lecture on mindfulness (control, n = 23). Subsequently, participants completed a Dutch-language empathic accuracy task. **Results:** There was no significant overall difference in empathic accuracy between the three participant subgroups, suggesting no effect of the two target interventions. Nonetheless, even though empathic accuracy appeared unaltered by facial emotion recognition training among participants who understood Dutch well, it was better after this intervention than after the control intervention among participants with a relatively limited understanding of Dutch. **Limitations:** The study used a small convenience sample. The control condition was listening to a lecture on mindfulness. Empathic accuracy was not assessed at baseline. Moreover, we did not formally assess language understanding, as we did not predict its presumed impact a priori. **Conclusions:** A better study design is needed to find out whether facial emotion recognition training can help improve empathic accuracy when the understanding of verbal cues is limited.

Keywords *emotion recognition, empathic accuracy, facial expressions, interpersonal, mindfulness*

[1]Department of Psychology, University of Groningen, The Netherlands

[1]Department of Psychology & School of Behavioural and Cognitive Neurosciences, University of Groningen, The Netherlands

Received
October 3, 2022
Accepted
September 21, 2023
Published
November 9, 2023
Issued
December 18, 2023

Correspondence
University of Groningen
m.aan.het.rot@rug.nl

One component of social interactions is empathy. Empathy is impaired in various mental disorders. Mindfulness interventions and social cognition training can enhance empathy over the course of weeks (Birnie et al., **2010**; Lam et al., **2011**; Mascaro et al., **2013**; Mazza et al., **2010**; Russell et al., **2006**; Russell et al., **2008**). To determine if this effect may also occur within shorter periods, we examined the acute impact of (a) a brief mindfulness exercise, and (b) basic facial emotion recognition (FER) training on empathic accuracy (EA).

Empathy and psychopathology

Empathy is considered a component of social cognition and can be broadly described as the capacity to understand the behaviour of others, to experience their feelings, and to express that understanding to them (Lam et al., **2011**). Affective empathy is concerned with one's emotional reactions to others' feeling states and cognitive empathy is the ability to recognize and identify these feeling states. Cognitive empathy is closely linked to Theory of Mind (ToM), which denotes the capacity to realize that others' minds and perspectives can differ from one's own (Cuff et al., **2016**).

One form of cognitive empathy found to be altered in the context of psychopathology is empathic accuracy (EA), defined as the ability to accurately infer others' feeling states (Ickes, **1997**) and operationalized in lab studies as the correspondence between the feel-

Take-home Message

In individuals presumably relying on non-verbal information to understand others' emotions, we found empathic accuracy to be higher after a brief facial emotion recognition training but not after a brief mindfulness exercise. However, we considered these results inconclusive because empathic accuracy and emotion understanding were not assessed at baseline.

ings reported by a target and the feelings a perceiver infers from the target's emotional expressions (Zaki et al., **2008**). As psychopathology is generally characterized by impairments in interpersonal functioning, and EA is considered key to effective social interactions (Ickes, **1997**), increasing EA in individuals with a mental disorder characterized by low EA might help improve their interpersonal functioning and thereby lessen their symptoms.

Psychological interventions can improve empathy over time (Birnie et al., **2010**; Lam et al., **2011**; Mascaro et al., **2013**). However, few studies have examined their immediate effects. In comparison, there have been studies on the acute impact of biological interventions. Specifically, EA can increase after one dose of oxytocin (Bartz et al., **2010**) and decrease after drinking alcohol (Thiel et al., **2018**). These experimental studies suggest EA to be malleable over short time periods. We aimed to add to these findings by examining the acute impact of two psychological interventions, mindfulness and FER training.

Mindfulness and empathic accuracy

While there is no universally accepted definition, mindfulness is often considered to reflect a non-judgmental awareness of the present moment (Bishop et al., **2004**). Mindfulness interventions generally aim to increase attention to this present moment, acceptance of thoughts and feelings, and self-awareness (Sauer-Zavala et al., **2013**).

Mindfulness interventions can increase empathy (Lam et al., **2011**). One potential mechanism of this effect is increased awareness of one's internal physical state (Birnie et al., **2010**; Fischer et al., **2017**; Sauer-Zavala et al., **2013**; Shapiro et al., **1998**). This interoceptive awareness is often trained by means of body-scan exercises. During these exercises, individuals attend to different body parts and the sensations they are experiencing in the present moment. Body-scans can have an immediate effect on state mindfulness (Upton & Renshaw, **2019**). Also, while their acute impact on interoceptive awareness remains unstudied, body-scans can increase interoceptive awareness over time (Fischer et al., **2017**).

As interoceptive awareness is considered a component of self-awareness, which itself is considered important for empathy (Gallup Jr & Platek, **2002**), body-scans might also increase empathy. To date, this effect of body-scan exercises on empathy remains unknown. No past empathy study has examined body-scans in isolation. Body-scans are part of the mindfulness-based stress reduction (MBSR) program by Jon Kabat-Zinn. While two MBSR studies have shown positive effects on self-reported empathy (Birnie et al., **2010**; Shapiro et al., **1998**), neither study specifically evaluated how body-scans contributed to these effects.

Another limitation of these two previous studies is their use of subjective empathy measures. While studies on the effects of mindfulness meditation, another MBSR component, have assessed empathy more objectively (e.g., Mascaro et al., **2013**), these studies' measures involved artificial social interactions or still images of facial expressions. Therefore, their generalizability to real life is considered limited.

In short, the acute effect of an isolated body-scan on a performance measure of empathy high in ecological validity has not been measured.

Facial emotion recognition (FER) and empathic accuracy

Emotions may be communicated both verbally and nonverbally. While verbal (auditory) communication appears more important than nonverbal (visual) communication, both contribute to EA (Zaki et al., **2009**). Facial expressions in particular are considered a crucial source of nonverbal information regarding others' feelings, particularly when others are more (rather than less) expressive and expressing negative

(rather than positive) feelings. Improving the ability to recognize how targets feel from their facial expressions may enhance perceivers' ability to interpret information about targets' emotional states. Indeed, teaching individuals the distinct features of facial expressions representing specific emotions can have this effect (Beitel et al., **2005**). In other words, FER training may help increase empathy.

Two intervention studies that involved social cognition training, including FER training, showed promising effects in individuals with schizophrenia. For example, emotion recognition and ToM improved after 12 weeks of Emotion and ToM Imitation Training (Mazza et al., **2010**). However, this training included not only FER but also mimicking facial expressions, inferring others' internal states from sketches, and assessing others' intentions from observing their actions. Consequently, the study only provides indirect evidence for the idea that FER training may increase empathy.

Another study found improved EA after a one-week isolated FER training (Russell et al., **2008**). This training used the Micro-Expression Training Tool (METT) developed by Paul Ekman, which includes short video-clips to teach the facial features of micro-expressions of emotion. A pilot study by the same group suggested that EA might even improve after a single session (Russell et al., **2006**). However, in both studies the EA measure was a simple emotion-matching task, with limited ecological validity. Also, participants were individuals with schizophrenia and matched controls; FER training may have different effects in other samples.

In short, the acute effect of a brief FER training on a performance measure of empathy considered high in ecological validity has not been measured.

The present study

We examined the acute effect of (a) a brief mindfulness exercise, namely a body-scan, and (b) basic FER training on EA. Similar to previous studies on the acute impact of oxytocin or alcohol on EA (Bartz et al., **2010**; Thiel et al., **2018**), we used a between-groups design. We hypothesized that EA would be higher among participants who completed either intervention than among participants who completed neither.

To assess EA we used the same task as Thiel et al. (**2018**). Participants are presented with a series of video-clips of targets talking about autobiographical emotional events and using a continuous rating dial to indicate how these targets were feeling while talking. This setup is thought to make the task highly ecologically valid. Participants simultaneously watch and listen to the targets as they share personal experiences from their actual lives.

We expected both interventions to be effective in acutely increasing EA. Participants assigned to the FER training would show improved task performance because we employed the METT, which teaches how emotions are featured on specific areas of the face. As such, the FER training was expected to promote other-awareness and thereby increase cognitive empathy, including EA.

Participants assigned to the body-scan were also expected to show improved performance on the EA task. This exercise can acutely increase state mindfulness (Upton & Renshaw, **2019**). By enhancing their awareness of their internal physical state, individuals may also become more emotionally aware and thereby show an increased capacity for affective empathy. As affective empathy can provide input during the process of understanding others (Cuff et al., **2016**), an increased capacity for affective empathy may lead to increased cognitive empathy, including EA. Overall, by comparing the effect of the body-scan and the FER training, we expected to learn more about the roles of the self and the other in obtaining EA, respectively, thereby highlighting its interpersonal nature (Zaki et al., **2008**).

Finally, while the language of the EA task was Dutch, participants in our study had a varied understanding of Dutch. We subsequently explored between-person variation in Dutch-language comprehension as a moderator of the effects of the two interventions on EA.

Method

Participants

We recruited sixty-nine participants (62% female) who were first-year students from the Dutch and English Psychology Bachelor programs at the University of Groningen. Their mean age was 20 years (*SD* = 2). Dutch was the mother tongue of 23 participants; 22 completed the questionnaires in Dutch and one in

English (who was in the English program). The remaining 46 participants had another mother tongue (46% German, 4% English, 16% other); 45 completed the questionnaires in English and one in Dutch (who understood Dutch fluently and had the Dutch nationality).

Measures

Baseline questionnaires

All participants provided basic demographic information and completed two Likert scales ranging from 0 (not at all) to 4 (very good) to assess their fluency in understanding Dutch and English, respectively (i.e., Dutch/English-language comprehension).

Trait mindfulness was assessed using the Five-Factor Mindfulness Questionnaire (FFMQ; Bohlmeijer et al., **2011**). It includes 24 statements rated from 1 to 5, with higher scores indicating greater mindfulness. The FFMQ previously demonstrated adequate to good internal consistency (Bohlmeijer et al., **2011**). However, in the present sample internal consistency of both language versions was poor (Cronbach coefficient α's of 0.23-0.46).

Trait empathy was assessed using Empathy Quotient (EQ; Groen et al., **2015**; Lawrence et al., **2004**). Respondents indicate their level of agreement with 40 statements (e.g., "I find it easy to put myself in somebody else's shoes"). Around half of the items are reversed to avoid response bias. The English EQ previously demonstrated good reliability and validity (Lawrence et al., **2004**). In contrast, psychometrics for a Dutch translation were previously shown to be better when 28 statements were used (Groen et al., **2015**). Consequently, we used a revised Dutch EQ including these 28 statements and 14 distractors. Both this version and the 40-item English EQ demonstrated acceptable internal consistency (α's of 0.78-0.79).

Outcome measure

EA was assessed using a Dutch-language task developed by aan het Rot and Hogenelst (**2014**) and programmed in E-Prime 2.0 (Psychology Software Tools). The original task includes two sets of 20 video-clips, in which female and male targets describe past personal experiences that are either positive (e.g., falling in love) or negative (e.g., a family member dying). The autobiographical nature of the clips makes the task high on ecological validity. Moreover, aan het aan het Rot and Hogenelst (**2014**) previously demonstrated that EA task performance can be predicted from scores on a validated empathy questionnaire. The present study used one of the two previously validated sets and, due to time constraints, 16 out of the 20 video-clips.

The clips lasted on average around 2 minutes. Clip selection was pseudo-randomized: all participants watched an equal number of positive and negative clips but never watched more than two clips of the same valence consecutively or the same target twice consecutively. While watching, participants were instructed to pay attention to both verbal and nonverbal cues and to continuously rate the emotional state of the target using a rating dial that corresponded to a Likert scale presented onscreen (1 = extremely negative, 9 = extremely positive).

Similarly, targets had previously provided continuous ratings of their own clips (aan het Rot & Hogenelst, **2014**). These self-ratings were used as reference for evaluating participants' performance. In line with previous work, for each clip, participants' and targets' continuous ratings were averaged across five-second intervals, the first and last intervals were discarded, and the remaining ratings were correlated, yielding scores between -1.00 and +1.00. These EA scores were subjected to Fisher's z transformation prior to data analysis.

Procedure

Upon arrival in the lab, students received written study information. The study's stated purpose was to examine the impact of attention training on how people perceive others' feelings. Any questions concerning the study were answered before participants signed consent forms.

Participants first completed the baseline questionnaires. Secondly, they were assigned to one of the interventions using block randomization and order of participation. Thirdly, they completed the Dutch EA task. Fourthly, they answered questions about the perceived difficulty of the procedures; their accuracy in

responding; and their ideas regarding the true study purpose. Before leaving the lab, participants were debriefed. Participation was compensated with partial course credit.

Each intervention lasted around 10 minutes. The mindfulness intervention involved listening to a recording of a guided body-scan developed by Elisha Goldstein. Doing the exercise while listening has previously shown to increase state mindful awareness by Ostafin & Vollbehr (unpublished work). The audio-clip directed participants to pay attention to their body parts while using their breath to stay in the present moment, and to adopt a non-judgmental, accepting attitude towards their experienced feelings and thoughts. The original video-clip is available at http://elishagoldstein.com/videos/10-minute-body-scan/.

For FER training we employed the Micro-Expression Training Tool (METT). Participants were presented with examples of facial expressions of happiness, sadness, anger, and fear, and instructed to direct their attention towards the associated muscle movements, which are the nonverbal cues conveying the particular emotion. More about the METT can be found at https://www.paulekman.com/micro-expressions-training-tools/.

Participants in the control condition listened to a lecture on mindfulness by Elisha Goldstein, specifically the part about the reasons for why mindfulness is not an inborn skill. This control was also previously used by (Ostafin & Vollbehr, unpublished work). The original video-clip is available at https://www.youtube.com/watch?v=bT-BCCkpmU7o/.

Data analysis

We used SAS 9.4 for Windows for all analyses. For significance testing the α was set at 0.05. Findings are reported using estimated least-squares means and standard errors (SE), unless indicated otherwise. Participant data (not target data to ensure confidentiality) and SAS syntax are freely available on DataverseNL: https://doi.org/10.34894/NLPJRL.

To examine baseline demographic and trait data, we used either general linear models with intervention (mindfulness, FER training, control) as the between-subjects factor or, for data with a nominal scale, X^2 tests. All subsequent analyses were done using hierarchical linear models with maximum likelihood estimation, following Kenward and Roger (**1997**) for computing the denominator degrees of freedom. Given previous results by aan het Rot and Hogenelst (**2014**), we first tested whether EA differed by (1) target gender, and (2) the valence of the video-clips. See models 1 and 2 in Table 2. There was a main effect for target gender, $F(1,68) = 4.20$, $p = 0.04$, with participants obtaining lower EA for male than for female targets. There was no significant main effect for valence, $F(1,68) = 0.22$, $p = 0.64$.

To test our hypothesis that participants who completed the body-scan or the FER training would score higher on EA than participants in the control condition, we first entered the main effect for intervention as predictor (model 3) and then the main effects for target gender and intervention as predictors (model 4). Follow-up analyses in case of a significant main effect for intervention are described below.

To explore whether the intervention effect on EA might be moderated by participants' level of Dutch-language comprehension, we entered the target gender, main effects for intervention and understanding Dutch, and the intervention by understanding Dutch interaction as predictors. Scores on understanding Dutch were grand-mean centred prior to analysis.

Effect sizes for each intervention effect are expressed as Cohen's *d* values.

Results

Baseline data

Understanding of English (used in the interventions) ranged from 2 to 4 ($M = 3.5$, $SD = 0.6$). Understanding of Dutch (used in the EA task) ranged from 0 to 4 ($M = 2.0$, $SD = 1.6$). As expected, participants whose mother tongue was Dutch understood Dutch better, $M = 4.0$, $SD = 0.0$, than participants whose mother tongue was not Dutch, $M = 1.0$, $SD = 0.8$, $t(45) = 25.79$, $p < 0.0001$. Both of these language subgroups understood English to a similar degree, $M = 3.4$, $SD = 0.6$, and $M = 3.5$, $SD = 0.6$, respectively, $t(67) = -1.02$, $p = 0.3$.

There were similar numbers of participants in each intervention subgroup (Table 1). There were no significant differences between the subgroups on any of the demographic and trait variables.

Hypothesis testing: Effect of the interventions on EA

The mean untransformed EA score (*r*) across all 1104 participant / video-clip combinations was 0.45 (range -1.00 to +1.00). Among Dutch participants, the mean untransformed EA score (*r*) was 0.61. Among non-Dutch participants, the mean untransformed EA score (*r*) was 0.37. Data analyses involved Fisher's *z* transformed scores, but untransformed scores are occasionally mentioned for interpretation purposes.

See Table 2 for multilevel regression analysis results. The main effect for intervention was not significant in model 3, $F(2,66) = 0.79$, $p = 0.46$, $d = 0.22$, nor in model 4, $F(2,66) = 0.79$, $p = 0.46$, $d = 0.22$. Further, when we added target gender as a moderator instead of as a covariate (model 5), this result did not change, and there was no significant intervention by target gender interaction, $F(2,66) = 1.71$, $p = 0.19$. Furthermore, when we examined the main effect for intervention for video-clips of male versus female targets separately, it was not significant for either target gender (male, model 3a: $F(2,66) = 2.06$, $p = 0.14$, $d = 0.35$; female, model 3b: $F(2,66) = 0.32$, $p = 0.73$, d = 0.14).

Moreover, when we added valence as a moderator instead of target gender (model 6), there was no significant intervention by valence interaction, $F(2,66) = 0.83$, $p = 0.44$. Indeed, when we repeated this analysis for clips of male versus female targets separately (models 6a-6b, this result did not change (effect for interaction with male targets: $F(2,66) = 0.01$, $p = 0.99$, $d = 0.02$; effect for interaction with female targets, $F(2,66) = 1.21$, $p = 0.30$, $d = 0.27$). In sum, hypothesis testing provided no evidence for an acute impact of the interventions on EA.

Exploratory analysis: Dutch-language comprehension as a moderator

We explored whether the intervention effect on EA might be moderated by participants' level of Dutch-language comprehension (model 7) because participants completed the EA task in Dutch yet varied in their understanding of Dutch. Participants who understood Dutch less were expected to perform worse on the EA task, thereby having more room for improvement.

The main effect for intervention was again not significant, $F(2,63) = 1.48$, $p = 0.23$, $d = 0.31$. However, the main effect for understanding Dutch, $F(1,63) = 54.32$, $p < 0.0001$, and the intervention by understanding Dutch interaction were significant, $F(2,63) = 3.73$, $p = 0.03$. Testing the interaction effect involved comparing the slopes for the different conditions. Among participants with a higher understanding of Dutch, the difference in slopes for FER training versus control was not significant, $b = 0.08$ (SE 0.14), $t(63) = 0.59$, $p = 0.56$, $d = 0.15$, indicating there were no significant differences in EA between these two conditions. Similarly, the difference in slopes for mindfulness versus control was not significant, $b = -0.01$ (SE 0.14), $t(63) = -0.08$, $p = 0.94$, $d = 0.02$, nor was the difference in slopes for FER training versus mindfulness, $b = -0.09$ (SE 0.15), $t(63) = -0.62$, $p = 0.54$, $d = 0.16$. Among participants with a lower understanding of Dutch, the difference in slopes for mindfulness versus control was also not significant, $b = -0.03$ (SE 0.15), $t(63) = 0.22$, $p = 0.83$, $d = 0.05$. However, the difference in slopes for FER training versus control was significant, $b = -0.40$ (SE 0.15), $t(63) = -2.73$, $p = 0.0082$, $d = 0.69$, as was the difference in slopes for FER training versus mindfulness, $b = 0.37$ (SE 0.14), $t(63) = 2.70$, $p = 0.0090$, $d = 0.68$.

Figure 1 visualizes the result of this follow-up analysis and was generated by computing point estimates for the transformed EA scores (Fisher *z*) at each level of condition and at two levels of understanding Dutch (higher versus lower, defined as 1 standard deviation above versus below the mean, respectively). Simple contrasts between the three interventions at these two levels of understanding Dutch conservatively used an adjusted α of 0.05 / 6 = 0.0083. Untransformed EA scores (*r*) averaged 0.56 after the mindfulness intervention, 0.56 after the FER training, and 0.58 in the control condition among participants with a higher understanding of Dutch, and 0.25 after the mindfulness intervention, 0.38 after the FER training, and 0.24 in the control condition among participants with a lower understanding of Dutch.

To ensure that this finding was not confounded by clip valence, model 8 also included this variable as a covariate, with results comparable to model 7.

Finally, as participants whose mother tongue

was not Dutch had a lower language understanding than participants whose mother tongue was Dutch, we repeated the analysis but used the dichotomous variable mother tongue (Dutch, other) instead of the continuous variable understanding Dutch. As expected, there was a main effect for mother tongue, $F(1,63) = 29.02$, $p < 0.0001$, which confirmed that the participants whose mother tongue was Dutch performed better on the EA task. However, neither the main effect for intervention, $F(2,63) = 0.27$, $p = 0.76$, $d = 0.13$, nor the intervention by mother tongue interaction were significant, $F(2,63) = 1.85$, $p = 0.16$. This suggests that FER training improved task performance in participants who would have otherwise performed poorly due to their limited understanding of the language of the task, rather than due to their mother tongue per se.

Discussion

To find out whether EA might be acutely malleable by a psychological manipulation (see Purpose), we examined the effect of a brief mindfulness exercise and basic FER training. We hypothesized that participants who completed either psychological intervention would obtain higher EA scores, assessed with a Dutch-language performance task, than participants who did not. However, the results of our hypothesis testing did not indicate that EA was improved by either intervention.

No immediate effect of increased mindfulness on EA?

Many past studies have reported positive effects of mindfulness interventions on empathy, including EA (e.g., Lam et al., **2011**). While some studies used self-report measures of empathy (Birnie et al., **2010**; Shapiro et al., **1998**), others used more objective measures (e.g., Mascaro et al., **2013**). Overall, while some studies have not found significant effects of mindfulness interventions on EA, there is consensus that they can improve empathy.

Nonetheless, we found that a 10-minute body-scan did not acutely improve EA. This was unexpected as previous studies have reported improved empathy following similarly brief mindfulness interventions. For example, Tan et al. (**2014**) found positive effects on both affective and cognitive empathy after a 5-minute breathing exercise, and Winning and Boag (**2015**) found increased cognitive empathy after a 15-minute mindfulness meditation, particularly in more extravert or conscientious participants. This suggests that the length of our intervention alone cannot explain the null result.

Instead, the type of intervention may account for this. While we used a body-scan to increase EA, Tan et al. (**2014**) used a breathing exercise and Winning and Boag (**2015**) used mindfulness meditation. Both studies checked that their intervention increased state mindfulness. Similarly, however, body-scans have previously been reported to increase state mindfulness (Upton & Renshaw, **2019**). This argues against the idea that the intervention type might help explain differences between our and previous results (but see Limitations below).

We had hypothesized that the body-scan exercise would improve EA by increasing interoceptive awareness (Fischer et al., **2017**), which is thought to contribute to emotional awareness which in turn is thought to be important for empathy (Cuff et al., **2016**; Gallup Jr & Platek, **2002**). However, the link between interoceptive awareness and emotional awareness may be less strong than we assumed. Indeed, while Sauer-Zavala et al. (**2013**) found improvements in self-awareness after three weekly body-scans, sitting meditation, or mindful yoga, the latter two interventions had larger effects than first one.

Possible impact of FER training on EA

Although the results of our hypothesis testing did not indicate that EA was improved by either the body-scan or the FER training, we additionally explored whether participants' understanding of Dutch could moderate the effect of both interventions. We found that among participants with a relatively limited understanding of Dutch, EA was higher in the subgroup who completed the FER training than in the subgroups who either completed the control condition or the body-scan, see Figure 1. Participants who understood Dutch well showed high EA task performance regardless of their assigned condition.

Participants whose understanding of Dutch was relatively limited presumably could not rely

on verbal auditory information (e.g., affective language) and may thus have focused primarily on nonverbal auditory information (e.g., affective prosody) and visual information (e.g., facial expressions). Participants who completed the FER training were explicitly instructed to examine facial expressions as nonverbal cues of emotional states. This suggests the FER training may have benefited participants with a limited understanding of Dutch because they became better perceivers of the visual information presented in the video-clips, i.e., of targets' emotional states. In other words, they may have benefited from the FER training thanks to improved visual emotion processing.

EA generally requires processing of both verbal and nonverbal emotion information. Experimental support for this idea comes from Zaki et al. (**2009**) who studied EA in English-speaking individuals using an English-language task but assigned some individuals to watching the video-clips without sound and others to listening to the video-clips without images. EA was lowest when only visual information was present, which underscores the importance of auditory (including verbal) information for EA. However, EA was also reduced when only auditory information was present, which shows that visual information also contributes to EA.

Our finding of increased EA after FER training in individuals whose understanding of Dutch was relatively limited similarly highlights the potential value of visual information when inferring others' emotional states. FER training may improve empathy by increasing the focus on visual information, particularly in interpersonal situations in which verbal information is not readily available. If so, then FER training might be particularly useful in individuals with auditory information processing impairments. This could include individuals with schizophrenia, which has previously been associated with low EA (Lee et al., **2011**). In line with this idea, previous studies have shown effects of FER training on other aspects of empathy in individuals with schizophrenia (Mazza et al., **2010**; Russell et al., **2006**; Russell et al., **2008**).

Overall, FER training might be more likely to benefit situations or individuals characterized by verbal understanding difficulties. In contrast, if verbal understanding is unaffected, FER training may be of little benefit. However, although both past and present findings are in line with this idea, the present findings are limited by multiple study limitations.

Limitations of the present study

One potential drawback of our study was its reliance on but a small sample of psychology students. The sample size limits interpretation of the statistically non-significant results. Psychology students tend to score high on self-report measures of empathy, for example when compared with students of the natural sciences (Thomson et al., **2015**). This might have increased the likelihood of a ceiling effect, at least among participants who understood Dutch well. However, their mean untransformed EA score (r) was 0.61, which is comparable to Thiel et al. (**2018**), who did not sample psychology students. Also, the maximum score is +1.00, which indicates that there was room for improvement.

As for the interventions, one drawback of our study is that they were offered in English, which was not the mother tongue of many participants. Consequently, some participants may have had difficulties in understanding the body-scan exercise or the METT, resulting in no increase in state mindfulness or FER in these participants. However, all participants understood English reasonably to very well, thereby reducing the chance that this had a significant impact.

Nonetheless, an additional shortcoming of the control condition may have been that listening to a lecture on mindfulness could actually have had a positive impact on participants' attitudes concerning mindfulness, thus increasing their emotional awareness. This effect might help explain the non-significant differences between the control condition and the two other conditions. Asking participants to listen to a neutral didactic lecture on a topic unrelated to mindfulness (or empathy) might prove to be a better control condition.

Similarly, one shortcoming of the two experimental conditions was that we did not include a manipulation check. Thus, while body-scans have previously been reported to increase state mindfulness (Upton & Renshaw, **2019**), we did not examine this. Similarly, while Emotion and ToM Imitation Training, which includes FER training, has previously been shown to improve FER accuracy (Mazza et al., **2010**),

we did not assess FER accuracy before and after our FER training. If our interventions did not have the intended effects on state mindfulness and FER accuracy, respectively, then this could also help explain the non-significant differences between the conditions.

Some uncertainty remains as to whether the findings were due to the interventions or some pre-existing group differences. In terms of our outcome measure, we note that the EA task was administered after the interventions but not before. A repeated-measures design would have allowed for a better test of the effectiveness of the interventions. While the task can be administered twice (aan het Rot & Hogenelst, **2014**), we did not do this due to time constraints.

As a final note, the internal consistency of the Dutch and English FFMQ was low. This is in line with increasing concerns about its cross-cultural validity (Medvedev et al., **2018**). Nonetheless, as we only used the FFMQ to assess trait mindfulness, its low internal consistency is immaterial for the outcome of our study.

Suggestions for future research

Though our study results are preliminary, they suggest that that FER training might be able to improve EA in participants whose ability to understand others is reduced due to a limited understanding of others' spoken language. As no previous study has examined the immediate effects of a brief FER training on EA, future research should aim to test this idea using a better study design.

Additionally, follow-up studies might help clarify the mechanisms by which FER training might increase EA. For example, to examine whether improved recognition of happy vs. sad expressions, shown in the positive vs. negative video-clips shown during the EA task, might contribute to increased EA after FER training, the "test" function of the METT could be utilized (as it assesses FER accuracy). This would provide information on the specificity of the FER training in terms of its psychological effects. Conversely, another way to assess this specificity would be to examine whether another type of training (e.g., language training, mindfulness training) would *not* improve FER accuracy.

A further avenue for follow-up studies could be to consider the different sources of information used for inferring others' emotional states during the EA task. Zaki et al. (**2009**) reported that English-speaking participants obtained higher scores on an English-language EA task when presenting only auditory information than when presenting only visual information. A future study in non-Dutch participants completing the Dutch-language task from the present study might find that participants only benefit from FER training when visual information is available, and not when only auditory information is available. This finding would confirm that FER training works by improving visual or facial emotion processing (and not by improving auditory or language information processing).

Conclusion

FER training might be of benefit to people aiming to visually infer the emotions of others in situations in which verbal cues are limited. This idea is relevant for future studies on how and when psychological interventions may increase EA. Importantly, the design of these studies should be carefully thought out, both in terms of how to experimentally test the impact of FER training and in terms of examining the role of verbal *vs.* non-verbal emotion understanding.

Funding

The authors have no funding to disclose.

Contributions

- **Marije aan het Rot** contributed to conception, design, data analysis and interpretation, and revising the article.
- **Merle-Marie Pittelkow** contributed to conception, design, data collection, data interpretation, drafting and revising the article.
- **D. Elisabeth Eckhart & Nils Simonson** contributed to conception, design, data collection, and revising the article.
- **Brian D. Ostafin** contributed to conception, design, data interpretation, and revising the article.

All authors gave final approval of the version to be published.

Compliance with Ethical Standards

All procedures of this study involving human participants were reviewed by the Ethics Committee of Psychology at the University of Groningen. The study was conducted in accordance with ethical standards comparable to the Declaration of Helsinki.

Conflicts of Interest

The authors declare they have no conflict of interest.

Informed Consent

All participants provided written informed consent.

Original Purpose

We previously studied the effects of alcohol administration on empathic accuracy. In the present study, we originally aimed to examine whether empathic accuracy might be acutely malleable by a psychological rather than a biological manipulation. The first psychological manipulation of interest was a mindfulness intervention, building on the idea that increasing self-awareness might lead to increased other-awareness. Thus, our initial hypothesis was that a brief mindfulness exercise could immediately improve empathic accuracy. The second psychological manipulation of interest was a facial emotion recognition (FER) training; this was also considered likely to increase other-awareness, and thus empathic accuracy, as participants were instructed on how to recognize others' emotions better. As a previous study successfully used a between-groups design to compare alcohol to placebo, we used a similar design in the present study. The idea to consider language understanding as a potential moderator evolved as the lead author was preparing the study for ethics review and realized we could study the role of language naturalistically in our intended sample: first-year Psychology students at our university complete their Bachelor program either in Dutch (the language of the empathic accuracy task) or in English, mostly depending on their mother tongue.

Acknowledgments

We wish to thank Sandra C. Krause for assistance with the study preparations.

References

aan het Rot, M., & Hogenelst, K. (2014). The influence of affective empathy and autism spectrum traits on empathic accuracy. *PLoS ONE*, *9*(6), 98436. https://doi.org/10.1371/journal.pone.0098436 (see pp. 88, 89, 93).

Bartz, J. A., Zaki, J., Bolger, N., Hollander, E., Ludwig, N. N., Kolevzon, A., & Ochsner, K. N. (2010). Oxytocin selectively improves empathic accuracy. *Psychological Science*, *21*(10), 1426–1428. https://doi.org/10.1177/0956797610383439 (see pp. 86, 87).

Beitel, M., Ferrer, E., & Cecero, J. J. (2005). Psychological mindedness and awareness of self and others. *Journal of Clinical Psychology*, *61*(6), 739–750. https://doi.org/10.1002/jclp.20095 (see p. 87).

Birnie, K., Speca, M., & Carlson, L. E. (2010). Exploring self-compassion and empathy in the context of mindfulness-based stress reduction (MBSR). *Stress Health*, *26*(5), 359–371. https://doi.org/10.1002/smi.1305 (see pp. 85, 86, 91).

Bishop, S. R., Lau, M., Shapiro, S., Carlson, L., Anderson, N. D., Carmody, J., Segal, Z. V., Abbey, S., Speca, M., Velting, D., & Devins, G. (2004). Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, *11*(3), 230–241. https:/ / doi.org/ 10.1093/ clipsy.bph077 (see p. 86).

Bohlmeijer, E. T., ten Klooster, P. M., Fledderus, M., Veehof, M. M., & Baer, R. (2011). Psychometric properties of the five facet mindfulness questionnaire in depressed adults and development of a short form. *Assessment*, *18*(3), 308–320. https://doi.org/10.1177/1073191111408231 (see p. 88).

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion Review*, *8*(2), 144–153. https://doi.org/10.1177/1754073914558466 (see pp. 85, 87, 91).

Fischer, D., Messner, M., & Pollatos, O. (2017). Improvement of interoceptive processes after an 8-week body scan intervention. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00452 (see pp. 86, 91).

Gallup Jr, G. G., & Platek, S. M. (2002). Cognitive empathy presupposes self-awareness: Evidence from phylogeny, ontogeny, neuropsychology, and mental illness. *Behavioral and Brain Sciences*, *25*(1), 36–37. https:/ / doi.org/ 10.1017/ S0140525X02380014 (see pp. 86, 91).

Groen, Y., Fuermaier, A. B., Den Heijer, A. E., Tucha, O., & Althaus, M. (2015). The empathy and systemizing quotient: The psychometric properties of the dutch version and a review of the cross-cultural stability. *Journal of Autism and Developmental Disorders*, *45*(9), 2848–2864. https://doi.org/10.1007/s10803-015-2448-z (see p. 88).

Ickes, W. (1997). *Empathic accuracy*. Guildford Press. (See pp. 85, 86).

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–997. https://www.ncbi.nlm.nih.gov/pubmed/9333350 (see p. 89).

Lam, T. M., Kolomitro, K., & Alamparambil, F. C. (2011). Empathy training: Methods, evaluation practices, and validity. *Journal of Multidisciplinary Evaluation*, *7*, 162–200 (see pp. 85, 86, 91).

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: Reliability and validity of the empathy quotient. *Psychological Medicine*, *34*(5), 911–919. https://doi.org/10.1017/s0033291703001624 (see p. 88).

Lee, J., Zaki, J., Harvey, P. O., Ochsner, K., & Green, M. F. (2011). Schizophrenia patients are impaired in empathic accuracy. *Psychological Medicine*, *41*(11), 2297–2304. https://doi.org/10.1017/S0033291711000614 (see p. 92).

Mascaro, J. S., Rilling, J. K., Tenzin Negi, L., & Raison, C. L. (2013). Compassion meditation enhances empathic accuracy and related neural activity. *Social Cognitive and Affective Neuroscience*, *8*(1), 48–55. https://doi.org/10.1093/scan/nss095 (see pp. 85, 86, 91).

Mazza, M., Lucci, G., Pacitti, F., Pino, M. C., Mariano, M., Casacchia, M., & Roncone, R. (2010). Could schizophrenic subjects improve their social cognition abilities only with observation and imitation of social situations? *Neuropsychol Rehabil*, *20*(5), 675–703. https://doi.org/10.1080/09602011.2010.486284 (see pp. 85, 87, 92).

Medvedev, O. N., Titkova, E. A., Siegert, R. J., Hwang, Y.-S., & Krägeloh, C. U. (2018). Evaluating short versions of the five facet mindfulness questionnaire using rasch analysis. *Mindfulness*, *9*(5), 1411–1422. https://doi.org/10.1007/s12671-017-0881-0 (see p. 93).

Russell, T. A., Chu, E., & Phillips, M. L. (2006). A pilot study to investigate the effectiveness of emotion recognition remediation in schizophrenia using the micro-expression training tool. *British Journal of Clinical Psychology*, *45*(4), 579–583. https://doi.org/10.1348/014466505X90866 (see pp. 85, 87, 92).

Russell, T. A., Green, M. J., Simpson, I., & Coltheart, M. (2008). Remediation of facial emotion perception in schizophrenia: Concomitant changes in visual attention. *Schizophrenia Research*, *103*(1-3), 248–256. https://doi.org/10.1016/j.schres.2008.04.033 (see pp. 85, 87, 92).

Sauer-Zavala, S. E., Walsh, E. C., Eisenlohr-Moul, T. A., & Lykins, E. L. B. (2013). Comparing mindfulness-based intervention strategies: Differential effects of sitting meditation, body scan, and mindful yoga. *Mindfulness*, *4*(4), 383–388. https://doi.org/10.1007/s12671-012-0139-9 (see pp. 86, 91).

Shapiro, S. L., Schwartz, G. E., & Bonner, G. (1998). Effects of mindfulness-based stress reduction on medical and premedical students. *Journal of Behavioral Medicine*, *21*(6), 581–599. https://doi.org/10.1023/A:1018700829825 (see pp. 86, 91).

Tan, L. B., Lo, B. C., & Macrae, C. N. (2014). Brief mindfulness meditation improves mental state attribution and empathizing. *PLoS ONE*, *9*(10), 110510. https://doi.org/10.1371/journal.pone.0110510 (see p. 91).

Thiel, F., Ostafin, B. D., Uppendahl, J. R., Wichmann, L. J., Schlosser, M., & aan het Rot, M. (2018). A moderate dose of alcohol selectively reduces empathic accuracy [journal article. *Psychopharmacology*, *235*(5), 1479–1486. https://doi.org/10.1007/s00213-018-4859-y (see pp. 86, 87, 92).

Thomson, N. D., Wurtzburg, S. J., & Centifanti, L. C. M. (2015). Empathy or science? empathy explains physical science enrollment for men and women. *Learning and Individual Differences*, *40*, 115–120. https://doi.org/10.1016/j.lindif.2015.04.003 (see p. 92).

Upton, S. R., & Renshaw, T. L. (2019). Immediate effects of the mindful body scan practice on risk-taking behavior [journal article. *Mindfulness*, *10*(1), 78–88. https://doi.org/10.1007/s12671-018-0948-6 (see pp. 86, 87, 91, 92).

Winning, A. P., & Boag, S. (2015). Does brief mindfulness training increase empathy? the role of personality. *Personality and Individual Differences*, *86*, 492–498. https://doi.org/10.1016/j.paid.2015.07.011 (see p. 91).

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science*, *19*(4), 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x (see pp. 86, 87).

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion*, *9*(4), 478–487. https://doi.org/10.1037/a0016551 (see pp. 86, 92, 93).

Tables

Table 1 Demographic, questionnaire, and task data for the three intervention groups

Demographic data	**Mindfulness (n=22)**	**FER training(n=23)**	**Control(n=24)**	**X^2/F**	**p**
Age in years	20 (2)	21 (3)	21 (3)	1.00	0.38
Female gender	64%	61%	63%	0.04	0.98
Dutch nationality	32%	39%	38%	0.29	0.87
Dutch as mother tongue	26%	35%	39%	0.57	0.75
Dutch language	32%	35%	33%	0.04	0.98
Questionnaire data[a]					
Understanding Dutch (range 0-4)	1.8 (2)	1.9 (2)	2.3 (1)	0.48	0.62
Understanding English (range 0-4)	3.7 (1)	3.3 (1)	3.5 (1)	1.95	0.15
FFMQ – Total score	71 (6)	69 (6)	70 (4)	0.82	0.44
EQ – Total score	38 (10)	43 (9)	43 (10)	1.74	0.18
Task data (post-intervention)[b]					
EA across film clips (Fisher's *z*)	0.7 (0.6)	0.9 (0.4)	0.8 (0.4)	0.75	0.47

Data expressed as mean (standard deviation) unless indicated otherwise. Gender was a binary variable. FFMQ = Five-Facet Mindfulness Questionnaire – Short Form. EQ = Empathy Quotient. EA = Empathic accuracy. Higher scores on understanding Dutch/English reflect a better Dutch/English-language comprehension. EA across film clips is expressed using Fisher's z transformed scores. (a) All questionnaires were administered in Dutch or English depending on whether participants were in the Dutch or English program of Psychology, respectively. (b) The EA task was in Dutch for all participants and administered post-intervention only.

Table 2 Results of multilevel regression analyses

Variables	Model 1	Model 2	Model 3	Model 3[a]	Model 3[b]	Model 4	Model 5	Model 6	Model 6[a]	Model 6[b]	Model 7	Model 8
Intercept	0.85 (0.07)***	0.88 (0.06)***	0.74 (0.10)***	0.46 (0.15)**	0.87 (0.09)***	0.79 (0.10)***	0.87 (0.11)***	0.73 (0.12)***	0.45 (0.18)*	0.89 (0.13)***	0.84 (0.08)***	0.87 (0.09)***
Level 1												
Valence(ref: Positive)												
Negative		-0.18 (0.09)*						0.03 (0.14)	0.03 (0.26)	-0.05 (0.17)		-0.06 (0.08)
Target sex (ref: Female)												
Male	-0.04 (0.08)					-0.18 (0.09)*	-0.41 (0.15)**				-0.18 (0.09)*	-0.18 (0.09)*
Level 2												
Condition(ref: : Mindfulness)												
Control			0.09 (0.14)	0.33 (0.21)	-0.03 (0.13)	0.09 (0.14)	-0.03 (0.15)	0.08 (0.17)	0.34 (0.25)	-0.08 (0.18)	-0.02 (0.10)	-0.02 (0.10)
FER training			0.17 (0.14)	0.40 (0.21)	0.07 (0.13)	0.17 (0.14)	0.07 (0.15)	0.28 (0.17)	0.39 (0.25)	0.21 (0.18)	0.14 (0.10)	0.14 (0.10)
Understanding Dutch											0.38 (0.07)***	0.38 (0.07)***
Understanding Dutch * Condition(ref: Mindfulness)												
Control											0.01 (0.10)	0.01 (0.10)
FER training											-0.23 (0.10)*	-0.23 (0.10)*
Level 1 * Level 2												
Target sex * Condition(ref: Female, Mindfulness)												
Male targets, control							0.35 (0.21)					
Male targets, FER training							0.32 (0.21)					
Valence * Condition(ref: Positive, Mindfulness)												
Negative, control								0.02 (0.19)	-0.03 (0.36)	0.10 (0.24)		
Negative, FER training								-0.21 (0.20)	0.02 (0.37)	-0.25 (0.24)		

(a) Only film clips involving male targets included in analysis. (b)Only film clips involving female targets included in analysis. Condition was coded 1 = control, 2 = FER training, 3 = mindfulness intervention. Target sex was coded 1 = male, 2 = female. Valence was coded 1 = negative 2 = positive. Understanding Dutch was a continuous variable; scores were grand-mean centred prior to analysis. Mother tongue was coded 0 = not Dutch, 1 = Dutch. Standard errors in parentheses; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Figures

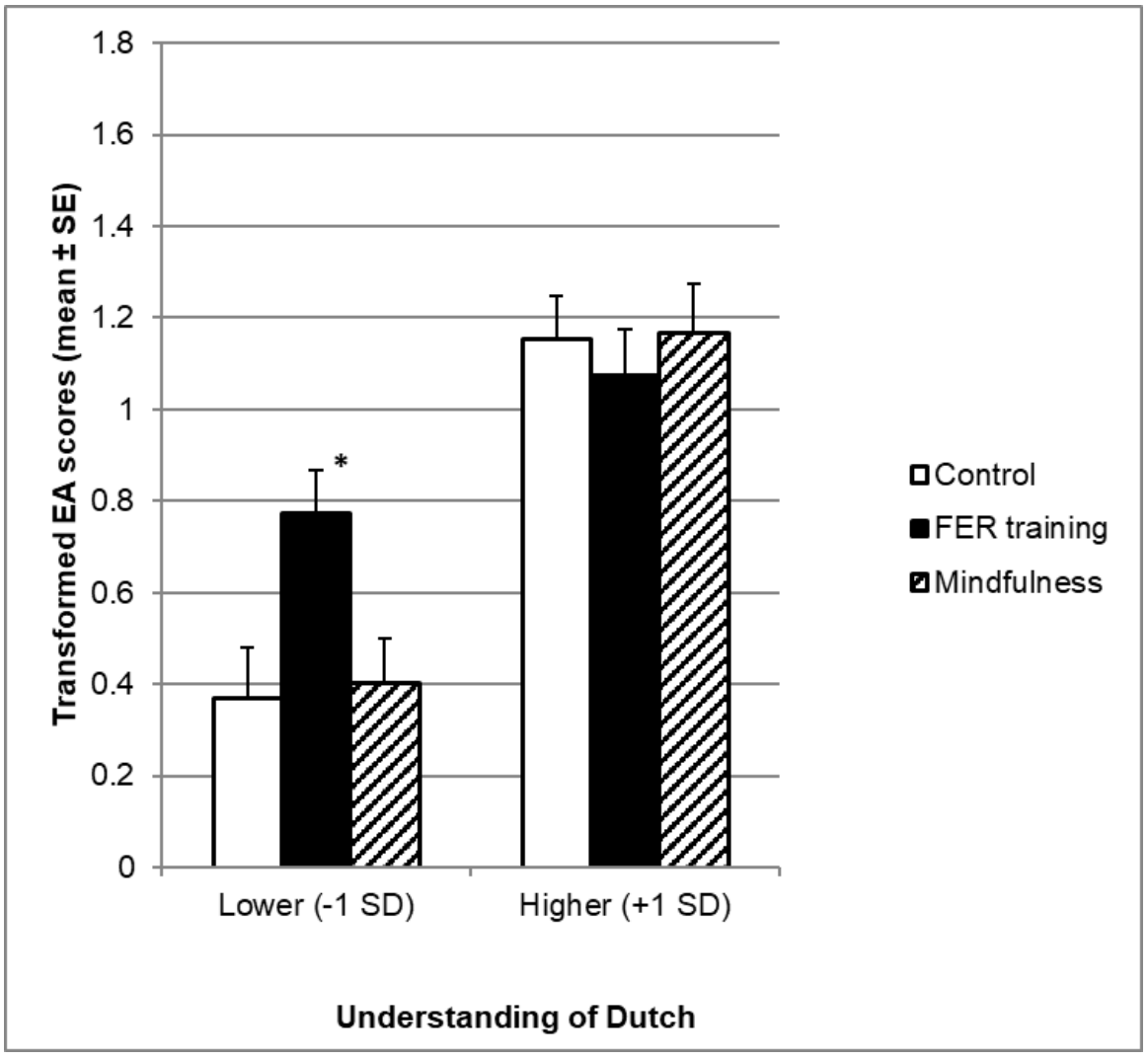


Figure 1 EA after intervention in participants varying in their understanding of Dutch. Note: *p<0.0083 (comparison with control intervention). EA = Empathic accuracy. FER = facial emotion recognition. SD = standard deviation. SE = standard error.

An Introduction to Complementary Explanation

Joeri van Hugten [1]

This paper introduces the practice of complementary explanation; the practice of taking a published result and writing a focused paper that rigorously and systematically describes the implications for a theory that would be rejected by those results. Such spotlighting of a rejected theory counteracts the common alignment between theory and result in published work.

Keywords *philosophy of science, falsification, publication bias, complementary explanation*

The reliability of the social sciences is threatened by underreporting. Underreporting refers to reported evidence not reflecting all collected evidence. This is concerning if reported evidence is a systematically disproportionate subset of collected evidence. Currently, this is the case with reported evidence being severely biased toward evidence that supports theories. For example, a large-scale replication of 100 psychology experiments replicated only 36 out of 97 significant results (Open Science Collaboration, **2015**). In industrial organizational psychology, hypotheses in journal articles are 73% supported and 12% rejected, while hypotheses in dissertations (which should be more representative of all collected evidence) are only 33% supported and 42% rejected (Mazzola & Deuling, **2013**; van Hugten & van Witteloostuijn, **2021**)

Underreporting is caused by underreporting practices such as hypothesizing after the results are known (HARKing) and not writing up all conducted tests. This leads to bias because especially results that do not support a paper's theory tend to be the ones not written up, and hypotheses made after the results are known tend to be ones that are in line with those results. Underreporting practices are prevalent. Measuring socially undesirable behavior is difficult, but best efforts suggest that 91% of academics know faculty who engaged in HARKing in the past year, 77% knows faculty who selected data that would support their hypothesis and withheld the rest (Bedeian et al., **2010**; Rubin, **2017**)

This paper proposes a practice to challenge theories and counteract underreporting. That practice is based on the falsificationist hypothetico-deductive philosophy (van Witteloostuijn, **2016**), because it opposes the beliefs underlying underreporting practices.

Underreporting practices as a neglect of falsification

Besides psychological factors (e.g., confirmation bias) and sociological factors (e.g., not undermining your colleague's theories), beliefs about what is important also underlie underreporting practices. In this section, I speculate about underlying beliefs for three aspects of underreporting practices, as well as a falsificationist principle that speaks to that belief. The posited beliefs are overlapping, and it turns out that falsificationist principles form a coherent opposition to those beliefs.

In the theory and hypothesis sections, why do HARKed hypotheses tend to be in line with the result? My personal intuition is that researchers understand that, at the level of the research program, theories aim to explain phenomena, but that this gets mistakenly transferred to believing that also hypotheses aim to explain results, at the level of the individual study. Given that aim, it follows that hypotheses that are not in line with the result are not useful (Johns, **2019**). By contrast, falsificationist principles maintain that hypotheses aim to challenge theory. For that aim, also hypotheses that are not in the line with the result can be

[1] Vrije Universiteit Amsterdam

Received
August 29. 2021
Accepted
May 12, 2022
Published
August 15, 2022
Issued
December 18, 2023

Correspondence
Vrije Universiteit Amsterdam
j.g.w.j.van.hugten@vu.nl

Table 1 Possible beliefs underlying underreporting practices.

Underreporting practice	*Possible underlying belief*	*Related falsificationist principle*
HARKed hypothesis tend to be in line with the result	Theories aim to explain the world, therefore hypotheses aim to explain the result	Hypotheses aim to challenge theory
Unreported results tend to be against the theory	Supported hypotheses give more valuable knowledge	Rejected hypotheses give more valuable knowledge
Discussion section focus on explaining the result	Explaining the result is the goal of a paper	Challenging the theory is the goal of a paper

Table 2 Complementary explanation in relation to existing practices.

	Focus on before the result is known	*Focus on after the result is known*
Focus on the result	Dominant practice	Abduction, CHarking, Harking, RHarking, Tharking
Focus on the rejected hypothesis	Counterargument, competing hypothesis, meaningful baseline, theory-driven null-hypothesis	Complementary explanation, SHarking

valuable (as long as they are tightly connected to a theory).

In the results section, why are results against hypothesis the underreported kinds of result? A possible underlying belief is that results that support hypotheses grant more valuable knowledge. In direct contrast, a key principle of falsification is that results against hypotheses give more valuable knowledge. For instance, seminal falsificationist Karl Popper argues that we learn more from results against hypotheses. Broadly speaking, the argument is that a result that supports a hypothesis does not imply that the theory is true, because that is affirming the consequent. Specifically, such an inference would go: 'if theory T is true, then data D should be observed' (i.e., the hypothesis), 'data D is observed' (i.e., the result is in line with the hypothesis), 'Therefore, theory T is true'. This is a logical fallacy because there may be alternative explanations for data D. Therefore, researchers try hard to rule out such alternative explanations by using random assignment, control variables, or more advanced statistical techniques. By contrast, a rejected hypothesis does imply that at least one premise in the theory or operationalization is false, because it is denying the consequent which is a valid form of argument (even if alternative explanations were not excluded). Less extremely, (Davis, **1971**) influentially argues that results that go against our expectations are more interesting.

In the discussion section, if a rejected hypothesis is reported, why is the expectation that authors explain the result? I speculate that the underlying belief may be that explaining data is a more important goal than improving theory. By contrast, falsificationist principles hold improving theory as the main goal. Therefore, those principles suggest that discussion sections build on the result to contribute contingencies that make the theory less simple or generalizable, and as a result, more accurate (e.g., Cross, **1982**; Lakatos, **1970**). Contributing contingencies can also happen in the process of explaining a result. However, the distinction is especially clear when discussion sections bring in a completely different theory that does fit the result. The distinction also becomes clearer if one imagines a more extreme alternate world in which discussion sections purposefully attempt to bring in additional theories that are opposite to the result. By contrast, current practice is that no further discussion is needed once the result is explained.

Overall, the argument is not that following the principles of falsification leads to more ethical research; it probably only affects the type of results that are underreported, not the extent of underreporting. That is, if researchers believed that rejected hypotheses lead to more valuable knowledge, then underreporting might start tending toward underreporting results that support hypotheses. Currently, the tendency is to underreport rejected hypotheses, so a practice based on principles of falsification can help bring balance.

A proposed counteracting practice: complementary explanation

Because of the opposition to falsificationist principles in the aspects of underreporting practices, I propose that a practice that

Table 3 CE Steps

Steps	Notes
0. Find a result	While reading, one might stumble upon a published finding that is striking if interpreted from the perspective of a different theory. The finding might even be merely a control variable for the original paper. Results with strong measures and research designs are ideal so that the result being opposite to a CE is clearly attributable to the theory.
1. Develop a CE for that result	What collection of premises suggest the opposite of the result? Premises that are straightforward and commonly held associations of concepts are ideal. That collection becomes the CE. If the result is opposite to the original hypothesis, then the original hypothesis development is a CE.
2. Identify a premise in that CE to challenge	By design, the CE is not in line with the result. So, at least one of its premises must be too simple.
3. Suggest a complication for that challenged premise	What would be one way in which we could complicate the challenged premise?
4. Evaluate that complication's effect on accuracy.	Does the complication increase accuracy? Is the complication plausible?
5. Iterate over steps 3-4. When out of ideas, compare complications.	The most simple and generalizable complication that can accurately predict the result is the ideal.
6. Iterate over steps 2-3-4-5. When out of ideas, compare challenges.	The less paper-specific the challenged premise, the greater the theoretical contribution.
7. Specify the contribution	Concisely and concretely describe the new insight. E.g., 'Premise 1 should be replaced by premise 1*' or 'Premise 1 is moderated by M'.

thoroughly applies those falsificationist principles can counteract underreporting practices. Specifically, I propose complementary explanation (CE).

The term 'complementary explanation' is a variation on the term 'alternative explanation'. An alternative explanation is an explanation for a result and an alternative to the hypothesis development (assuming that the result was in line with that hypothesis). Alternative explanations are the main threat that Popper aimed to avoid. By contrast, a complementary explanation (CE – countable) is an explanation for the opposite of a result, so it is a logical complement to the hypothesis development (assuming that the result was in line with that hypothesis). For example, if a quantitative study finds a positive coefficient, a CE for that result is a set of arguments that imply a negative coefficient. Similarly, for a qualitative study's causal story between high X and high Y, a theory's implication of a negative relation is a CE. If a study's result is inconsistent with its hypothesis, then the original hypothesis development is a CE. Even if a study does not have a hypothesis for a particular relation, an explanation of the opposite of its result is a CE. One result can have multiple CEs.

To appreciate CE's unique focus, Table 2 positions CE in the context of a comprehensive list of similar existing practices. CE is similar to counterarguments, competing hypotheses, meaningful baselines, or theory-driven null hypotheses (e.g., Schwab & Starbuck, **2012**). The difference is that CE is to be used after the result is known. Even more extremely, CE can be done after publication by someone who was not the original author.

CE is like HARKing and spinoffs like Tharking (i.e., transparently hypothesizing after the results are known (Hollenbeck & Wright, **2017**; Rubin, **2017**) and abduction (Locke et al., **2008**; Schwab & Starbuck, **2017**) in that all those practices happen after a result is found. However, the difference is that those practices aim to explain a result (although RHarking is, in principal, also open to rejected hypotheses (Rubin, **2017**)). For example, abduction would never involve explaining the opposite of the result. In other words, hypotheses made after the results are known tend to be ones that are in line with those results. But they need not be that way. CE is like transparently making a hypothesis after the result is known, that is opposite to that result. That shift in focus counteracts the threat of HARKing to research reliability. Finally, CE is like SHarking (suppressing hypotheses after the results are known); the most threatening form of HARKing (Rubin, **2017**), except that SHarking focuses on suppressing rejected hypotheses while CE adds exactly such hypotheses.

CE Steps

The steps to interpret supportive results seem clear: e.g., 1) $p<0.05$, 2) hypothesis supported, and 3) more confidence in the theory (but see Wasserstein et al. (**2019**) for how it is not that simple). By contrast, the application of falsification is impeded by a lack of such clear steps. CE is a way to codify falsificationist interpretation steps. Table 3 summarizes these steps.

A crucial step in falsification is that a rejected hypothesis implies that at least one premise in its explanation is false. But, it is undetermined exactly which one is false (Hines, **1988**; Lakatos, **1970**; Søberg, **2005**). That underdetermination can make falsification seem infeasible in practice. CE tackles this issue by evaluating theories based on a combination of their accuracy, simplicity, and generalizability (Weick, **1999**). If a result is against a theory, that means the theory has low accuracy. Then, we can trade off generalizability or simplicity for accuracy. For example, a trade-off for generalizability involves saying that the theory does not apply to the context of that result, and the theory will be accurate in contexts where it does apply. Alternatively, we sacrifice simplicity to restore accuracy, if we claim that the inconsistent result is due to a moderating contingency and once that moderator is considered, the result will be consistent with the theory.

The fact that such trade-offs are possible to 'save a theory from falsification' has been used to argue against falsification (Søberg, **2005**). Instead, CE views explicit discussion of such trade-offs as theory development. CE helps identify inaccuracies and make explicit what trade-offs are forced upon the theory. Theories (or research programs) with many such trade-offs are degenerate (Lakatos, **1970**). CE prompts and documents such degeneration. CE is not about the next step of judging whether degeneracy is significant enough. Potential users of a theory can judge whether the theory lost too much simplicity or generalizability to be useful. For example, see Cross (**1982**) judging monetarism (a research program in macroeconomics) while explicitly reflecting on the Lakatosian ideas at the basis of that judgment.

Step 0 and step 1 contribute by identifying a lack of accuracy. Step 0 may seem difficult, but the same creativity that is displayed in thinking of alternative explanations should also allow us to reinterpret results from theories that oppose that result. Regarding step 1, developing a CE does not require fully fleshed-out theories. Instead, CEs consist of the most straightforward, and commonly held, associations of concepts (i.e. accepted propositions in Davis, **1971**). The role of the following steps is to specify which association to make less straightforward; this is the complication where the theory's simplicity or generalizability is sacrificed for accuracy. That process of complication leads to the theory becoming more fleshed out rather than that a fully fleshed-out theory is required in step 1. Still, CEs in step 1 must have a level of explicitness, detail, and connection to literature more similar to hypothesis development than to the generally weak argumentation for alternative explanations (Spector & Brannick, **2011**).

Steps 2 to 6 contribute by identifying ways to restore accuracy by trading-off simplicity and/or generalizability. Thus, one of the accepted propositions is negated and replaced by a proposition that is more complex and more 'interesting' (Davis, **1971**). There may be cases where you have an intuition that a theory implies the opposite of a result, so you have found a CE, but then upon further reflection the implication is not so straightforward. For example, a gravitational theory predicts the location of a planet, but a result shows that the planet is not at that location. While writing the CE you discover that your intuition was simplistic; the theory only predicted that location under the assumption that there was no other nearby planet pulling the focal planet away from its orbit. In that case, it is tempting to scratch the CE. However, CE values making explicit this step of further reflection; showing the reader where the intuition needs to be complicated. Thus, a CE author may decide that many readers would have the same intuition, so explaining that complication is a valuable contribution. As another example, step 5 prompts CE authors to also present the second and third best challenges they came up with. For instance, maybe you challenged a premise by adding the complication that that premise only applies to gaseous planets (and the result was found for a solid planet). CE encourages including that challenge, even if another challenge ends up being more plausible.

Step 7 makes summarizes the complication; making explicit the degeneration that is forced upon the theory by the result. It is possible that a result is inconsistent with a CE because of bad measures, auxiliary premises, or research designs. The CE author can decide whether a CE with a step 7 that reads 'Measure X does not capture concept A (in some context)' contributes enough to be worth the effort. If a challenged premise is paper-specific, the contribution may be small. On the other hand, the contribution may be large enough if measure X is typical. For example, see Cook et al. (**1979**) discussing why insignificant results regarding the cognitive bias 'sleeper effect' are due to operationalizations not appreciating theoretical nuances. Cook et al. (**1979**)'s paper is like a CE paper, except that CE reframes the discussion from 'in this paper we remind people of some important nuances in the theory, which empirical studies have failed to appreciate, which led to insignificant results' to 'insignificant results have forced us to appreciate the importance of some nuances, and in this paper, we make explicit the nuances we now believe to be important'.

Full circle

Given those details, we can see how CE counteracts underreporting by improving meta-analyses. Meta-analyses use concept labels as inclusion criteria. For example, Heugens and Lander (**2009**)'s meta-analysis on 'mimetic pressure's effect on isomorphism' uses a variety of concept labels to search for literature (e.g., 'isomorphism', 'institutional theory'). Underreporting practices cause studies to be described in terms of concepts that are supported by the result. Therefore, meta-analyses disproportionately include studies that support the theory (Murphy & Aguinis, **2019**). (Note also that while Tharking (Hollenbeck & Wright, **2017**) and abduction (Locke et al., **2008**; Schwab & Starbuck, **2017**) do not mislead like HARKing, they still lead studies to be described in terms of concepts that are supported by the results).

Enter CE. For example, a study finds a positive effect of 'competition' (measured as the number of firms in the same industry) on 'differentiation'. A CE for that result is that a greater number of firms in the same industry can be interpreted as mimetic pressure (e.g., Haveman, **1993**) and differentiation is the opposite of isomorphism. Institutional theory suggests that mimetic pressure should increase isomorphism. Therefore, mimetic isomorphism theory implies a negative effect of the number of firms in an industry and differentiation; i.e., the opposite of the finding. Before the CE, this paper 'about differentiation' would fall outside of Heugens and Lander's meta-analysis inclusion criteria, so it would (systematically) fail to include results like these opposite to the theory. By contrast, after the CE is published, the result is described using theories that are not supported by it, so meta-analyses would include it. HARKing and not writing up tests could continue at the usual rate, but with CE, the proportion of rejected hypotheses among reported evidence would be greater (and closer to the true proportion).

Conclusion

Proposals to combat underreporting focus on preventing underreporting practices; e.g., deemphasize p-values (Bettis, **2012**), stop using "p<0.05", (Bettis, **2012**; Wasserstein et al., **2019**), or abandoning null-hypothesis significance testing (Schwab et al., **2011**). Such proposals are less feasible due to the inertia of current practice. By contrast, CE counteracts underreporting without preventing practices that lead to underreporting; it adds to, rather than changes, current practice. That increases feasibility. That is why the explanation of CE's value can assume that HARKing and not writing up tests continue at the usual rate. CE's value does not depend on whether a hypothesis was truly made after the results were known, nor does its value depend on what caused underreporting. CE makes use of the potential for interacting with studies after publication.

Moreover, CE is a useful practice, even if underreporting did not exist. First, CE also increases research reliability more directly. When CE is done by others than those who found the result, research reliability is increased simply by having an extra person thinking through the meaning of the data from a fresh perspective. Second, we put strain on others when collecting data. This comes with a responsibility to make the most of our data. CE helps fulfill that responsibility by reusing published results, in contrast to demands for efficient rather than comprehensive presentation, and novel findings.

In sum, I hope people use CE to learn more from the same findings and especially learn about, and from, those things that we currently miss due to underreporting.

Acknowledgements

I would to thank the reviewers, Pablo Martin de Holan, Jana Retkowsky, Arjen van Witteloostuijn, and the OT reading group at Tilburg University for their encouragement and valuable feedback on an earlier version of this work.

References

Bedeian, A., Taylor, S., & Miller, A. (2010). Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning & Education*, *9*(4), 715–725 (see p. 98).

Bettis, R. A. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, *33*(1), 108–113. https://doi.org/10.1002/smj.975 (see p. 102).

Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin*, *86*(4), 662–679. https://doi.org/10.1037/0033-2909.86.4.662 (see p. 102).

Cross, R. (1982). The Duhem-Quine Thesis, Lakatos and the Appraisal of Theories in Macroeconomics. *The Economic Journal*, *92*(366), 320. https://doi.org/10.2307/2232443 (see pp. 99, 101).

Davis, M. S. (1971). That's Interesting!: Towards a Phenomenology of Sociology and a Sociology of Phenomenology. *Philosophy of the Social Sciences*, *1*(2), 309–344. https://doi.org/10.1177/004839317100100211 (see pp. 99, 101).

Haveman, H. A. (1993). Follow the Leader: Mimetic Isomorphism and Entry Into New Markets. *Administrative Science Quarterly*, *38*(4), 593. https://doi.org/10.2307/2393338 (see p. 102).

Heugens, P. P. M. A. R., & Lander, M. W. (2009). Structure! Agency! (And Other Quarrels): A Meta-Analysis Of Institutional Theories Of Organization. *Academy of Management Journal*, *52*(1), 61–85. https://doi.org/10.5465/amj.2009.36461835 (see p. 102).

Hines, R. (1988). Popper's methodology of falsificationism and accounting research. *The Accounting Review*, *63*(4), 657–662 (see p. 101).

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, Sharking, and Tharking: Making the Case *for* Post Hoc Analysis of Scientific Data. *Journal of Management*, *43*(1), 5–18. https://doi.org/10.1177/0149206316679487 (see pp. 100, 102).

Johns, G. (2019). GUIDEPOST: Departures from conventional wisdom: Where's the next opposite effect? Academy of Management Discoveries. https://doi.org/10.5465/amd.2019.0226. (see p. 98).

Lakatos, I. (1970, September 2). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (1st ed., pp. 91–196). Cambridge University Press. https://doi.org/10.1017/CBO9781139171434.009 (see pp. 99, 101).

Locke, K., Golden-Biddle, K., & Feldman, M. S. (2008). Perspective—Making Doubt Generative: Rethinking the Role of Doubt in the Research Process. *Organization Science*, *19*(6), 907–918. https://doi.org/10.1287/orsc.1080.0398 (see pp. 100, 102).

Mazzola, J. J., & Deuling, J. K. (2013). Forgetting What We Learned as Graduate Students: HARKing and Selective Outcome Reporting in I–O Journal Articles. *Industrial and Organizational Psychology*, *6*(3), 279–284. https://doi.org/10.1111/iops.12049 (see p. 98).

Murphy, K. R., & Aguinis, H. (2019). HARKing: How Badly Can Cherry-Picking and Question Trolling Produce Bias in Published Results? *Journal of Business and Psychology*, *34*(1), 1–17. https://doi.org/10.1007/s10869-017-9524-7 (see p. 102).

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, *349*(6251) (see p. 98).

Rubin, M. (2017). When Does HARKing Hurt? Identifying When Different Types of Undisclosed Post Hoc Hypothesizing Harm Scientific Progress. *Review of General Psychology*, *21*(4), 308–320. https://doi.org/10.1037/gpr0000128 (see pp. 98, 100).

Schwab, A., & Starbuck, W. (2012). Using baseline models to improve theories about emerging markets. In C. Wang, D. Ketchen, & D. Bergh (Eds.), *West meets east: Toward methodological exchange (research methodology in strategy and management* (pp. 3–33, Vol. 7). Emerald Group Publishing Limited. (See p. 100).

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). PERSPECTIVE—Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests. *Organization Science*, *22*(4), 1105–1120. https://doi.org/10.1287/orsc.1100.0557 (see p. 102).

Schwab, A., & Starbuck, W. H. (2017). A Call for Openness in Research Reporting: How to Turn Covert Practices Into Helpful Tools. *Academy of Management Learning & Education*, *16*(1), 125–141. https://doi.org/10.5465/amle.2016.0039 (see pp. 100, 102).

Søberg, M. (2005). The Duhem Quine thesis and

experimental economics: A reinterpretation. *Journal of Economic Methodology*, *12*(4), 581–597. https://doi.org/10.1080/13501780500343680 (see p. 101).

Spector, P. E., & Brannick, M. T. (2011). Methodological Urban Legends: The Misuse of Statistical Control Variables. *Organizational Research Methods*, *14*(2), 287–305. https://doi.org/10.1177/1094428110369842 (see p. 101).

van Hugten, J., & van Witteloostuijn, A. (2021). The state of the art of hypothesis testing in the social sciences. In H. Mandele & A. van Witteloostuijn (Eds.), *A future for economics* (1st, pp. 167–185). VU University Press. (See p. 98).

van Witteloostuijn, A. (2016). What happened to Popperian falsification? Publishing neutral and negative findings: Moving away from biased publication practices (A. Klarsfeld, L. C. Ng, & S. Eddy, Eds.). *Cross Cultural & Strategic Management*, *23*(3), 481–508. https://doi.org/10.1108/CCSM-03-2016-0084 (see p. 98).

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "$p < 0.05$". *The American Statistician*, *73*, 1–19. https://doi.org/10.1080/00031305.2019.1583913 (see pp. 101, 102).

Weick, K. (1999). Conclusion: Theory construction as disciplined reflexivity: Tradeoffs in the 90s. *The Academy of Management Review*, *24*(4), 797–806 (see p. 101).