

This is only a PDF export as local copy. The official preregistration can be found under:

<https://osf.io/tsb68/>

## Testing the effectiveness of a lateral reading intervention on source trustworthiness discernment among German Internet users

### Study Information

#### *Hypotheses*

H1s: Lateral reading improves source trustworthiness discernment compared to control.  
H1c: Lateral reading improves claim credibility discernment compared to control. H2s: Lateral reading improves source trustworthiness discernment compared to control two weeks after the treatment. H2c: Lateral reading improves claim credibility discernment compared to control two weeks after the treatment. H3c: Online search reduces claim credibility discernment compared to control (i.e., backfire effect).

### Design Plan

#### *Study type*

Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

#### *Blinding*

For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

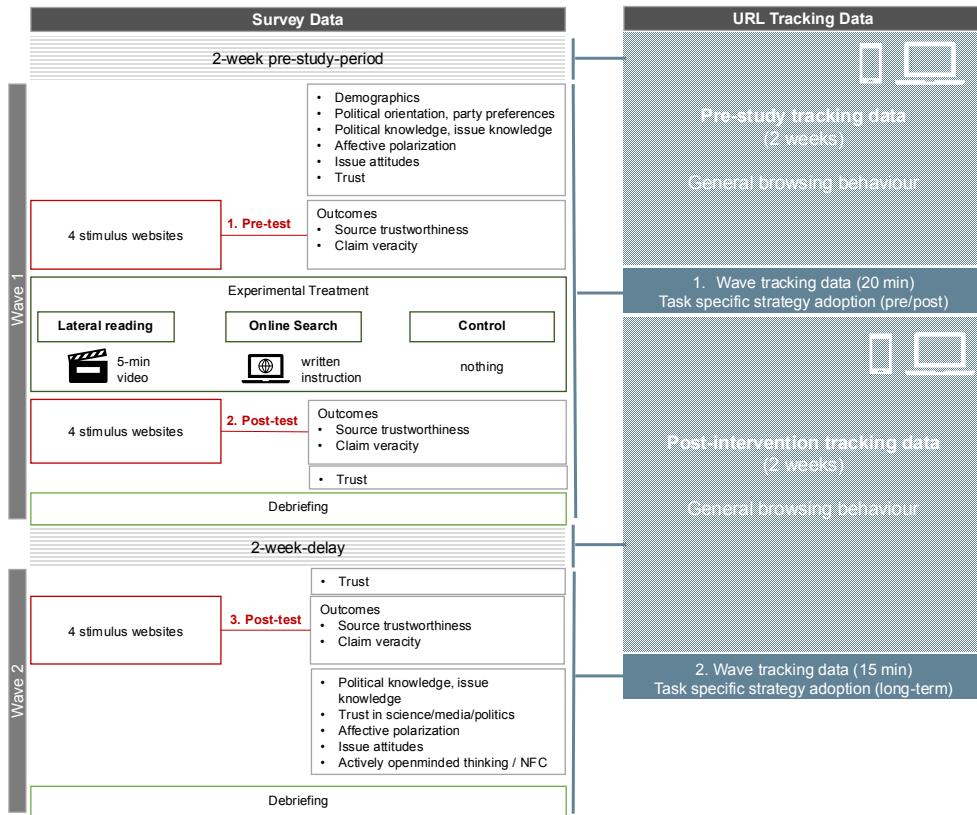
Personnel who interact directly with the study subjects (either human or non-human subjects) will not be aware of the assigned treatments. (Commonly known as “double blind”)

#### *Is there any additional blinding in this study?*

No data

#### *Study design*

We have a between subjects design with 1 factor and 3 levels (lateral reading, online search, control) with two waves.



### *Randomization*

Participants are randomly assigned to one of the three experimental conditions (lateral reading, online search, control) with equal probability (1/3). Participants are individually randomised at the start of the survey and those who replace drop outs will also be individually randomised. This randomisation allocation happens after YouGov confirms participants' ability to watch a video. If participants can't watch a video (mostly for technical reasons) they are screened out and are replaced before the random allocation happens. The selection of stimuli is fully randomized between participants and across waves with the constraint that in each phase (pre, post, follow-up) 2 trustworthy/true items and two untrustworthy/false items will be presented.

### Sampling Plan

### *Existing Data*

Registration prior to creation of data

### *Explanation of existing data*

No data

### *Data collection procedures*

Survey data and URL tracking data will be collected by YouGov. The study will be fielded to the German URL tracking panel (YouGov Pulse) that tracks ca. 750 individuals on mobile and desktop devices. To complement the tracking sample, the survey will be fielded to a representative sample of ca. 2000 individuals without tracking who can take the survey on mobile or desktop devices. For the main analysis of treatment effects, samples from both panels will be pooled. Participants are compensated by YouGov. According to their internal policies, participants who drop out before completing the first wave of the study will not be compensated. Participants who fail their internal attention checks but complete the survey, will be compensated. Excluded participants (due to drop-out or inattentiveness) will be replaced by other participants in the data to reach the agreed sample size of complete submissions. Participants who complete the first wave but fail to complete the second wave will be compensated partially. Participants are remunerated via YouGov's internal point system.

*No files selected*

### *Sample size*

Constrained by our research budget, we will recruit 750 participants via the YouGov Pulse panel as well as further 2000 participants via YouGov's main panel, where we will specify our recruitment criteria to include a balanced representation of the German public. YouGov will replace any participants who fail to complete a survey or do not pass their internal attention check, thus ensuring we reach the agreed sample size of completed survey registrations.

### *Sample size rationale*

The sample size is fully determined by our research budget.

### *Stopping rule*

*No data*

Variables

### *Manipulated variables*

Experimental condition 1 - Lateral Reading The treatment for the lateral reading condition is delivered in the form of a 5-minute-long video explaining the strategy of lateral reading and demonstrating an example examination of an untrustworthy website with a screen recording. The video is presented after the first set of 4 websites. Experimental condition 2 - Online Search The instructions to search online to find evidence regarding the central claim were directly translated from Aslett et al. (2023) and presented after the first set of 4 websites (parallel to the lateral reading video). Control The control condition receives neither the lateral reading video, nor the online search instructions but continues with the rating of websites without treatment.

*No files selected*

### *Measured variables*

The stimulus material consists of a set of 6 trustworthy and 6 untrustworthy websites containing online articles regarding a specific political domain (Covid-19, climate change, migration, international organizations, elections or the war in Ukraine). 4 websites will be selected for the pre-treatment assessment, another 4 will be selected for the post-treatment assessment and the remaining 4 will be shown to participants in the second wave of the survey as follow-up post-treatment test. In all conditions, the instructions to rate the outcomes are held constant. Furthermore, in all conditions, a 2-minute time limit (per

website) is mentioned in the instructions to limit the time participants spend on the website assessment (it is not enforced technically but only in the instructions). The instructions are the following (all translated from German): General Instruction (presented once) In the following you will visit different websites. We will then ask you for your personal evaluation of the website. As you probably know, you often come across advertising on the Internet. Sometimes you will be asked to consent to direct messages, sign up for a newsletter, or even make a donation. Please do not respond to any of these requests when visiting the websites within our study. Some of the following websites contain untrustworthy information. At the end of the study, we will inform you in detail which websites are not trustworthy and why. Instruction for every website (presented for every website) "Please follow the link [link to trustworthy or untrustworthy website with article]. The website should open in a new browser tab. Please take up to 2 minutes to reach your judgment." Outcomes for every website Source trustworthiness How trustworthy do you find the website? (definitely trustworthy, probably trustworthy, probably not trustworthy, definitely not trustworthy) Source intent What do you think is the intent of the website? (inform, convince, mislead) Claim credibility How credible do you find the central claim of the article? (definitely credible, probably credible, probably not credible, definitely not credible) Claim veracity If you had to make a clear decision, do you think the central claim is true or false? (true, misleading and/or false, could not determine) In addition to the main outcome variables, participants will rate their familiarity with the source. Survey Measures In addition to the main outcome variables, several demographic, political and psychological variables are collected as covariates and/or moderating variables. Demographics We collect information on participants' age, gender, education, and employment status. Political Variables We collect information on participants' political orientation, party preferences, voting intentions, affective polarization, populist attitudes, political extremism, political knowledge (efficacy), issue attitudes and issue knowledge and institutional trust. Psychological variables We include the construct of "active open-minded thinking" (AOT) including items on myside-bias and overconfidence.

*No files selected*

### *Indices*

There are no indices/scores required to test our hypotheses.

*No files selected*

### Analysis Plan

#### *Statistical models*

Our primary analyses examine the changes in trustworthiness discernment ability (difference between trustworthiness ratings for trustworthy and trustworthiness ratings for untrustworthy sources), as well as the changes in claim credibility discernment (difference between credibility ratings of true vs. false articles) across two timeframes: (a) pre-intervention versus post-intervention (immediate treatment effect), and (b) pre-intervention versus 2-week follow-up (longer-term treatment effect). Source trustworthiness ratings serve as the first outcome (H1s, H2s). The analysis will be repeated with claim credibility ratings (H1c, H2c, H3c). As additional robustness check, source intent and binary claim veracity measures will be explored and reported supplementary. We will employ robust linear mixed-effects models, implemented in R using the `robustlmm::rlmer()` function which robustly fits linear mixed-effects models using the Robust Scoring Equations estimator.

Models 1s and 1c (pre vs. post), to test hypotheses H1s, H1c and H3c, will model the rating using the equation:

rating ~ item\_type \* phase \* treatment + (1 + item\_type + phase | id) + (1 + phase + treatment | item)

- rating refers to (model 1s, H1s) the source trustworthiness rating assigned to an item or (model 1c, H1c) the claim credibility assigned to an item;
- treatment is a factor variable representing the three conditions;
- item\_type is a factor variable indicating whether the sources in the item are trustworthy or untrustworthy, thus quantifying discernment ability (i.e., the difference between mean trust judgments in trustworthy sources and mean trust judgments in untrustworthy sources);
- phase is a factor variable distinguishing between pre- vs. post-intervention judgments;
- id serves as an anonymous unique identifier for each participant; and
- item is an unique identifier for each item.

Models 2s and 2c (pre vs. post. vs. follow-up), to test hypotheses H2s and H2c, are identical to models 1s and 1c except: (a) only participants who have completed the follow-up wave will be included, and (b) the variable phase will now consist of the levels pre vs. follow-up instead of pre vs. post. For each of the models described earlier, we will select a model specification by initially implementing the full random effect structure for both participants and items (see the above equations). If necessary, following Bates et al. (2014 <https://arxiv.org/abs/1506.04967>) we will iteratively simplify the random effect structure (e.g. to address issues of singular fit). If this does not resolve the issues, we will use lme4::lmer() with the full random effect structure (and, if necessary, again simplify the random effect structure). The experimental treatment will be dummy coded with the control group specified as reference category. We compare both treatment groups (lateral reading and online search) to the control group.

*No files selected*

### *Transformations*

No data transformations are required to test our core hypotheses. For exploratory analyses, continuous variables will be scaled by subtracting the mean and dividing by two standard deviations as suggested by Gelman (2008 <https://doi.org/10.1002/sim.3107>).

### *Inference criteria*

We use one-tailed tests and report fixed effects coefficients for the interaction term with 90% confidence intervals (applying the conventional alpha-threshold of 0.05).

### *Data exclusion*

There will be no further exclusions other than those made internally at YouGov for survey drop-out and failing internal attention checks.

### *Missing data*

We do not expect missing values on our core outcomes as YouGov requests complete responses from their panel members.

### *Exploratory analysis*

In addition to the primary analyses outlined above, we will also consider conducting the following supplementary investigations. Regarding the core experiment

- Investigating the extent to which the treatments influence participants' overall trustworthiness rating of the sources (pre vs. post & pre vs. follow-up), irrespective of their actual trustworthiness (i.e., the interaction `phase \* treatment` in Models 1 and 2; general skepticism effect).
- Analyzing changes in trustworthiness ratings separately for trustworthy and untrustworthy sources. In particular, we are interested in determining the degree to which the treatments cause participants to exhibit reduced trust in trustworthy sources.
- Determining whether participants' self-reported familiarity with a source (which we will elicit for each item) predicts trustworthiness ratings (i.e., familiar sources are trusted more). For this purpose, we would explore extending Models 1 and 2 with interactions involving a dummy variable that indicates whether a participant reported encountering a website before participating in the study.
- Verifying whether the immediate treatment effects (pre vs. post) in Model 1 (all participants) and Model 2 (only participants who completed the 2-week follow-up) are comparable. Exploring the robustness of results with binary claim veracity and source intent as supplementary outcome measures. Involving URL tracking data
- Investigating whether participants' use of lateral reading (measured through tracking data and as self-reports) increases in the lateral reading condition. We assess whether participants' use of lateral reading moderates the observed treatment effects (H1s). To achieve this, we will consider extending Models 1 and 2 with interactions involving a dummy variable that indicates whether a participant used lateral reading. These models can also be estimated as LATE (local average treatment effect among compliers - those who use lateral reading) while the core models follow an ITT (intent to treat) approach.
- Investigating whether lateral reading reduces the downstream prevalence of untrustworthy sources in participants' media diet, measured through tracking data (2 weeks pre-treatment browsing vs. 2 weeks post-treatment browsing).
- Examining whether treatment effects are moderated by the prevalence of untrustworthy sources in participants' media diets. More specifically, we determine whether the interventions are effective among the subgroup with high prevalence of untrustworthy websites in their media diet. We report different cut-offs for high prevalence. Involving further variables
- Investigating whether lateral reading has downstream consequences for political variables such as affective polarization, political issue attitudes and political trust.
- Regarding the covariate of media trust, investigating whether the overall study task of evaluating websites causes a loss in media trust (across all conditions and split between conditions). We examine whether the entire study task, due to unusually high exposure to untrustworthy content, causes short-term reduction in media trust (comparing pre vs. post evaluation task measures of media trust in both waves) and whether this effect decays shortly after, not causing a long-term loss of media trust (comparing the post-evaluation measure of the first wave to the pre-evaluation measure of the second wave, two weeks later).
- Investigate the extent to which intervention effects vary depending on the following factors: education, age, party preferences, and political orientation, populist attitudes and actively open-minded thinking.

## Other

### *Other*

Information on URL Tracking Data Research questions involving URL tracking data are not formally preregistered due to the complexity of analytical decisions. However, below we outline how we envision a pre-processing and data analysis pipeline could look like. Raw URL tracking data contains the following variables: full URL of visited website, timestamp of

visit, duration of visit in seconds. Preprocessing. While some analyses will be performed on the full URLs (e.g. topic matching, Google search term analysis), other analyses will be conducted on the domain level (e.g. NewsGuard classification). When collapsing the data to the domain level, we will consider the number of clicks, as well as the duration of visits. The top 1000 domains will be classified manually (and potentially using existing domain lists), to filter out, for our study, clearly irrelevant website visits (e.g. banking, porn, etc.). We follow best practices for preprocessing (von Hohenberg, 2023) and publishing (Munzert et al., 2023) of sensitive URL tracking data (in combination with survey data). General description. We classify all potentially relevant domains for which labels are available using NewsGuard. Using those measures, we will create descriptive summaries of people's engagement with trustworthy and untrustworthy websites – in total and regarding different domains. We will then compare the measures of trustworthy media diets between conditions, as well as between the 2-week pre-treatment browsing period and the 2-week post-treatment browsing period. General topical browsing behavior. We will create a topic dictionary, containing words relevant for our stimuli domains (e.g. regarding climate change, health, etc.). When creating this dictionary, we will include keywords that have a semantic connection to the domain but we will also include relevant keywords that appear on SERPs when searching the stimuli domains. We will then apply the dictionary to the full URLs to identify websites potentially related to our stimuli websites. We will then construct chains of sequential website visits for individuals that occur in one context (one topical domain), and eventually create a network of in that regard linked websites across the entire sample (and between conditions). Ignoring this network structure, we will also consider the plain domain diversity within a topical domain. Lateral Reading technique adoption. As a measure of lateral reading technique adoption during the experimental website assessment, we will consider whether people "googled" the stimuli websites by extracting Google search terms from the full URLs. We will compare the extent to which this behavior occurs between conditions. As a second measure of technique adoption, we will also compare the lengths of subsequent website visit chains, within one topical domain, between conditions, during the time of the experiment. Ignoring this network structure, we will also consider the plain domain diversity within a topical domain, within the time of the experiment. We will also consider the time spent within and beyond the stimulus websites during the experiment. Stimulus Material Stimulus websites for this German sample were selected using a systematic bottom-up strategy involving an external data provider (NewsGuard), a systematic protocol and a pre-test with N = 301 participants on Prolific for stimulus selection. Articles that were uninformative for the assessment of participants' veracity and trustworthiness discernment ability, for example due to ceiling effects, were disregarded from the stimulus pool for the main study. The stimulus pretest was registered under the Open Science Framework prior to data collection and received ethical approval independently from the main study.