Preregistration

Videos in peer-review of Registered Reports

Lisa Spitzer[1], Tobias Heycke[2]

[1] Leibniz Institute for Psychology Information

[2] GESIS - Leibniz-Institute for the Social Sciences

14. July 2020

Hypotheses

Predictions

1) We expect that participants that can also watch a screen recording of an experimental procedure in addition to reading about it ("text + screen recording condition") identify more errors that were implemented into a Registered Report Protocol, than participants that can only read about the procedure ("text condition"), when requested to review the procedure.

2) We expect a main effect of condition in multiple rating scales:

- 2a) First, we expect participants in the "text + screen recording condition" to rate the procedure to be *less* complex, as compared to the rating of participants in the "text condition".
- 2b) We expect participants in the "text + screen recording condition" to indicate a *better* understanding of the procedure, as compared to participants in the "text condition".

- 2c) We expect participants in the "text + screen recording condition" to indicate a *lower* cognitive load, as compared to participants in the "text condition".

Rationale (optional)

We base our hypotheses on the argumentation of Heycke & Spitzer (2019) who argued that sharing screen recordings might increase the reproducibility of computer assisted data collection procedures, because small details that are not mentioned in the final manuscript (but might be important for measuring the tested effect) are visible in the video. Furthermore, it was argued here that videos represent an easy way to show an experimental procedure, because they can easily be accessed without the need for running any (possibly proprietary) software.

In this paper, it was also contemplated that screen recordings might not only be beneficial for conducting replication attempts, but also could improve the peer review process. Peer review requires many resources that reviewers often do not have (e.g. because they have to manage reviews in addition to their normal workload), because deeply understanding a described procedure requires much effort. Screen recordings could provide reviewers with a faster and deeper understanding of the procedures by showing the procedure in a quick and intuitive way. Furthermore, it might be the case that errors or flaws of the procedure might be more easily detected if reviewers can also watch a screen recording, for example because the procedue cannot be described as detailed in words as it is visible in a video, and flaws are thus more visible in a video. Investigating these possibilities is the aim of this study.

Methods

Design

To test our hypotheses, we will use a 2 (condition: "text condition" vs. "text + screen recording condition") x 3 (manuscript: manuscript A based on Heycke & Stahl (2018) vs. manuscript B based on Schneider, Stapels, Koole, & Schwarz (2020) vs. manuscript C based on Rosemann & Thiel (2018)) between-subjects design.

Planned sample

We have received funding for our study from the Fetzer Franklin Fund of the John E. Fetzer Memorial Trust. Based on the available funding, we will collect data of $N = 102$ participants. We have conducted a sensitivity analysis to investigate what

effect sizes can be detected based on this sample. This yielded a detectable effect size of $d = 0.66$ for our main analysis (hypothesis 1), using a t test for independent samples with $\alpha = \beta = .05$. Even though this effect size is rather large, it seems reasonable to expect a large effect for the condition difference ("text condition" vs. "text + screen recording condition") regarding the number of detected errors.

- t tests - Means: Difference between two independent means (two groups)
- Analysis: Sensitivity: Compute required effect size
- Input: Tail(s) = One
  - α err prob = 0.05
  - Power ($1 - \beta$ err prob) = 0.95
  - Sample size group 1 = 51
    - Sample size group 2 = 51
- Output: Noncentrality parameter $\delta = 3.3122423$
  - Critical $t = 1.6602343$
  - Df = 100
  - Effect size $d = 0.6559217$

To ensure that all manuscripts are represented equally and data collection is stopped when the maximum of paid participants is reached, $n = 34$ participants will be collected in each manuscript condition (as $34 * 3 = 102$).

Exclusion criteria

We will screen-out participants at the beginning of our study, that do not belong to our target sample (they will be re-directed to an exit page instead of the main study). Specifically, as we want to address psychological researchers, we will exclude all participants that indicate that they are not working in research, or that indicate that their research does not fall within the scope of psychology. We will only include data of participants that have completed all pages of the survey in our analyses. Furthermore, we will check for faithful participation at the end of the survey, and will only include participants into our analyses that indicate such faithful participation. This could lead to a smaller N than determined by the funding, as this exclusion is only made when the data is inspected after data collecton.

Material

The authors of the three used published papers (Heycke & Stahl, 2018; Rosemann & Thiel, 2018; Schneider et al., 2020) were contacted to request permission to use their papers for our study. Only introduction and methods section of these papers were used, as we wanted to imitate a "Registered Reports Protocol" which is submitted and peer reviewed prior to study administration. We decided to use this publication format, because we expect that the inclusion of screen recordings can be especially beneficial for peer review that is administered before the study, as flaws and errors can still be improved. We edited the manuscripts by adding ten errors/flaws and shortening them. A table listing all errors we included is available as part of this preregistration, to increase transparency and to document exactly what errors we will code later for our analyses.

Besides editing the manuscripts, we also created screen recordings based on the edited procedures (thus, also containing the errors). These will be displayed alongside the manuscripts in our study ("text + screen recording condition").

Procedure

The study was created with Soscisurvey (Leiner, 2019) and will be distributed via https://www.soscisurvey.de. It will take participants approximately 30-60 minutes to complete the study. When clicking on the study link, participants will be randomly grouped in one of the three manuscript conditions (we included multiple manuscripts to increase external validity). Additionally, they are also randomly assigned to the "text condition" or the "text + screen recording condition". Participants will be shown a welcoming page, and informed consent will be obtained. It will be inquired if participants are working in research, and if participants' research falls within the scope of psychology (and what topic they focus on, e.g., neuropsychology, general psychology). Additionally, they will be asked what academic group they belong to, what is their gender, and how many reviews they have written before. Then, participants will receive information about their task: They are instructed that they will read a Registered Report Protocol (consisting of introduction and methods section), and are asked to review the described procedure. Participants of the "text + screen recording condition" are instructed to also include the displayed screen recording into their review. To ensure a more standardized approach to the review, a *review guideline* is presented to participants prior to starting the task. In this guideline, it is described what errors/flaws can occur. As we used the differentiation

between “major errors” and “minor errors” to define the errors for our study, this differentiation is described in detail. This approach is based on prior research on the evaluation of peer review (Baxt, Waeckerle, Berlin, & Callaham, 1998).

Then, the main task will start: Participants will either just see the introduction and methods section, or additionally will also be able to watch a screen recording (either displaying the whole procedure, for the shorter experiments; or displaying the most relevant details, for the longer experiments). At the bottom of the page, they are asked to write down any flaws or errors they spotted.

After this review task, three rating scales will be displayed, asking about subjective understanding, an evaluation of the procedure's complexity, and cognitive effort. These scales are taken from a prior study conducted by the authors, in which a significant difference between “text condition” and “text + screen recording condition” was found for subjective understanding. Since these results were based on exploratory analyses, the scales will be examined again in the present study.

Lastly, participants are asked for comments, if they took any breaks (this will possibly be taken into consideration in exploratory analyses), and if they participated faithfully. Participants will then be asked for their e-mail address (which is collected separately) to receive their gift card (either for Amazon, Just Spices, or Thalia), which will be worth 15 €.

Two screen recordings of our procedure (for the “text condition” and “text + screen recording condition” of one manuscript) are shared as part of this preregistration. Screen recordings for the other conditions will be added when the study is submitted for publication.

Analysis plan

Confirmatory analyses

To gain higher power, all manuscripts will be pooled, thus the factor manuscript will not be regarded in the confirmatory analyses (yet, it might be analyzed on an exploratory basis later on).

To test hypothesis 1, we will code for each participant which of our a priori defined errors they have detected in their review. Only our pre-defined errors will be included for the analyses. We will use a one-sided Welch t test to compare the number of

errors detected by participants of the "text condition" vs. participants of the "text + screen recording condition".

To test hypothesis 2, we will run individual one-sided t tests to compare each scale between conditions ("text condition" vs. "text + screen recording condition"). The correction for multiple tests is described below.

We will furthermore calculate Bayesian t tests with Cauchy priors as addition to the frequentist t tests (scaling parameter of $r = \sqrt{2}/2$). For all t tests we also report the median of the posterior distribution as an effect size estimate and its 95% highest density interval (HDI).

Details of analysis

Missing data

We will only include data from participants that completed all pages, and responses to all confirmatory analyses are mandatory, therefore there will be no missing data for our confirmatory analyses.

Correction for multiple tests

As we will use three tests to test one set of hypotheses, we will use a Bonferroni-Holmes-corrections: $\alpha = .05/(3-1) = .025$

The following assumptions will be tested for our t tests: - Normality: - In case of $n > 30$ in each condition, a violation of this assumption is not problematic - If n is smaller, normality will be tested with Q-Q-plots and histograms using the R package `MVN` (Korkmaz, Goksuluk, & Zararsiz, 2014)

- Independence of samples:
  - Is achieved via the experimental design
- Independence within each sample:
  - Is achieved via the experimental design

Existing data	**Data collection has not begun.**
Project schedule (optional)	Based on our funding requirements, data will be collected beginning on July 15, 2020, and ending either at the end of August (so data will be collected for 1 1/2 months), or until the limit of paid participants is reached. Data will be analyzed between September and October, and the manuscript will be written and submitted by the end of the year.

References

Baxt, W. G., Waeckerle, J. F., Berlin, J. A., & Callaham, M. L. (1998). Who reviews the reviewers? Feasibility of using a fictitious manuscript to evaluate peer reviewer performance. *Annals of Emergency Medicine*, *32*(3), 310–317. doi:10.1016/s0196-0644(98)70006-x

Heycke, T., & Spitzer, L. (2019). Screen recordings as a tool to document computer assisted data collection procedures. *Psychologica Belgica*, *59*(1), 269–280. doi:10.5334/pb.490

Heycke, T., & Stahl, C. (2018). No evaluative conditioning effects with briefly presented stimuli. *Psychological Research*. doi:10.1007/s00426-018-1109-1

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An r package for assessing multivariate normality. *The R Journal*, *6*(2), 151–162. Retrieved from https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Leiner, D. J. (2019). SoSci survey. Retrieved from https://www.soscisurvey.de

Rosemann, S., & Thiel, C. M. (2018). Audio-visual speech processing in age-related hearing loss: Stronger integration and increased frontal lobe recruitment. *NeuroImage*, *175*, 425–437. doi:10.1016/j.neuroimage.2018.04.023

Schneider, I. K., Stapels, J., Koole, S. L., & Schwarz, N. (2020). Too close to call: Spatial distance between options influences choice difficulty. *Journal of Experimental Social Psychology*, *87*, 103939. doi:10.1016/j.jesp.2019.103939