Preregistration Report for Conceptual Replication of Experiment 1 by Wühr and Seegelke (2018)

Study Information

1. **Title**
Conceptual Replication of Experiment 1 by Wühr and Segeelke (2018)

2. **Authors**
Pia Stermann, Paul Jänsch, Ludger Masch, Alexander Prochnow

3. **Description**
The study is a conceptual replication of an experiment 1 from Wühr and Segeelke (2018) "Compatibility between Physical Stimulus Size and Left-right Responses: Small is Left and Large is Right". We investigate stimulus-size-horizontal-response mapping compatibility effects by means of an online experiment. For example, Wühr and Segeelke (2018) found that participants in their experiment 1 were faster to respond to a large stimulus object with a key that is on the right of the keyboard (e.g. "Backspace", pressed with the right hand), than with a key that is on the left of the keyboard (e.g. "Tabulator", pressed with the left hand). For small stimulus objects, they found that reaction times were numerically faster for left-hand responses than for right-hand responses. The background of this study is "A theory of magnitude" (ATOM), which was proposed by Walsh (2003). The theory basically claims that there exist certain relations between the cortical representations of "time, space and quantity". Many studies reported evidence for this theory in the form of compatibility effects between numerical size and horizontal response location (Dehaene et. al., 1990), size-congruity effects between numerical size and physical size (e.g., Henrik & Tzelgov, 1982) and compatibility effects of stimulus size and horizontal response location (Ren et al. ,2011; extended by Wühr and Segeelke, 2018).

4. **Hypotheses**

   Hypotheses (1)-(3) will be tested for right-handed participants only.

   (1) Response times for right-hand responses are faster to the larger stimulus than to the smaller stimulus.

(2) Response times for left-hand responses are faster to the smaller stimulus than to the larger stimulus.

(3) The stimulus size – response location compatibility effect is larger for right-hand responses than for left-hand responses. This means, the difference in response times for right-hand responses to be faster to the larger stimulus than to the smaller stimulus is larger than the difference in response times for left-hand responses to be faster to the smaller stimulus than to the larger stimulus.

Additionally, we will test the following hypotheses for left-handed participants, to investigate whether we can find a compatibility effect in opposite direction for left-handed people (opposing to Hypotheses (1)-(3)):

(4) Response times for left-hand responses are faster to the larger stimulus than to the smaller stimulus.

(5) Response times for right-hand responses are faster to the smaller stimulus than to the larger stimulus.

(6) The stimulus size – response location compatibility effect is larger for left-hand responses than for right-hand responses. This means, the difference in response times for left-hand responses to be faster to the larger stimulus than to the smaller stimulus is larger than the difference in response times for right-hand responses to be faster to the smaller stimulus than to the larger stimulus.

Finally, we will compare the size of possible compatibility effects regarding people's handedness under the following hypothesis:

(7) The stimulus size- response location compatibility effect is larger for right-handed people than for left-handed people. I.e., the absolute difference between response times in the compatible condition and response times in the incompatible condition is larger for right-handed than for left-handed people.

Design Plan

5. **Study type**
   Online experiment

6. **Blinding**
   (There is no explicit blinding involved in this study.) The relevant experimental manipulation of the dependent variable (i.e. the stimulus-response mapping) happens within-participants. Participants are neither informed about the purpose of their tasks in the experiment nor about the research hypotheses, but they are only presented with relevant information containing practical instructions on their task. This implies, that participants are informed about the change of stimulus-response mapping, since this is relevant for their task. However, they are not provided with any explanation for this change. The experiment is conducted via the internet. No direct contact between experimenters and participants will take place.

7. **Is there any additional blinding in this study?** No

8. **Study design**
The experiment uses a within-subject design with one factor: stimulus-response (S-R) mapping. The factor has two levels: the compatible condition (small stimulus=left-hand response and large stimulus=right-hand response), and incompatible condition (small stimulus = right-hand response and large stimulus = left-hand response). Our study consists of two main experimental blocks and one intermediate (non-experimental/ distraction) block. In the experimental blocks, which are further divided into training trials and experimental trials, a simple black square will be presented in each trial. This black square is either small or large, the actual displayed size of the stimuli depends on the screen size though the large square's edge length is always twice as long as the small square's edge length. Participants have to decide via a keypress (either "q" or "p") whether the square is small or large. In the compatible condition they have to press the "q" key with their left hand (index finger) when they see a small square and press the "p" key with their right hand (index finger) when they see a large square. In the incompatible condition, participants have to press the "q" key with their left hand (index finger) when they see a large square and press the "p" key with their right hand (index finger) when they see a small square. An equal number of trials for both S-R mappings (10 training and 60 experimental trials per condition) will be presented to all participants. The order of S-R mapping conditions is determined randomly, see next chapter "Randomization". The intermediate distraction task block is intended to distract participants from the initial S-R mapping and to limit training effects. The stimulus object in this block is either a simple black circle or a simple black triangle. The participants have to select with a mouse click one of two buttons under the stimulus to decide whether they see a circle or triangle. Further information on the detailed design, materials and procedure can be found in the attached document [experimentalDesign](#), accessible via our github repository.

9. **Randomization**
We will randomize the order in which each participant completes the experimental trial blocks by randomly choosing the initial S-R mapping condition for the first training and main block. The S-R Mapping condition of the following individual experimental trial blocks is determined by the (randomly chosen) initial S-R mapping (see the attached design document), such that each participant completes in total the same number of trials for each of the two mapping conditions (factor levels). The training and main experimental blocks each contain a fixed number of trials, in which one of the two stimulus items (namely large or small square) is presented, likewise the intermediate block with the "circle" or "triangle" stimulus. Within the trials of a block, the number of small and large stimuli (in the main experimental and training blocks) or circles and triangles (in the intermediate block) shown in total is equal, meaning that all participants see the same number of instances of either of the two stimulus objects within each block. The order in which these instances are presented is determined randomly and ad hoc, thus the sequence of small and large stimuli and the order of the S-R mapping in the experimental trials might vary (randomly) between participants and between trial blocks.

10. **Existing data**
Registration prior to creation of data: The data have not yet been created, since the experiment was not yet conducted. However, we collected data in a pilot experiment from N=4 participants. We used the format of this data (accessible here) to aid in the writing of our analysis script, but we did not pay attention to the outcomes of our analyses and will not include this data in our final analysis. The pilot mainly served to check for technical and comprehension issues. Additionally, when analyzing the pilot data and data from the original experiment, we recognized that a student-t distribution for the basis of our model leads to a significantly better fit of the data than a Gaussian (which we assumed earlier) when using leave-one-out cross-validation to compare the model's performance. Data from the previous study (N=24) by Wühr and Segeelke (2018) was available and guided the specification of statistical models. This data will not be included in the final analysis. No data from the experiment to be preregistered here was available at the time of preregistration.

11. **Explanation of existing data**
We are aware of the effect found in the original study and have reanalyzed the data created by this study. We will use in our eyes more fitting analyses on our collected data to test for a compatibility effect as reported in the original study. The data from our pilot study was only used to evaluate model fits and test our analysis script, but we ignored any statistical findings. For the corresponding analysis scripts for the original experiment click here and for the pilot click here.

12. **Data collection procedures**
Participants are advised to take part only once, however, we did not control for this. After having sent the initial invitations through social media and email (31st July 2020), we will wait until 3rd August 2020, 11:59 pm before closing data collection.
By following the link to the experiment, participants declare to have normal or corrected-to-normal visual acuity and should be naive with respect to the purpose of the study. Participation is voluntary and not rewarded with material things or money but the experimenters' deep gratitude.

13. **Sample size**
We aim to recruit as many participants as possible.

14. **Sample size rationale**
We conducted a power analysis to get an estimate for the minimal number of participants required to most likely accept our hypotheses in the case they are actually true. Since we have no former experience in performing such an analysis, the expressiveness of the result is limited if not even non-existent. For reasons of computational complexity, we only tested the power of one of our models, namely model_right which is responsible for assessing hypothesis 1 in particular.

We started by pre-compiling the model without drawing any samples. This was useful because we then were able to update the model with several different datasets without having to recompile the model each time which saved us a huge amount of time. Speaking of different datasets, we needed to simulate fake reaction time data for the compatible and incompatible mapping with which we then could supply our model. Because this data should be as realistic as possible for the experiment at hand, we used the mean and sd for both mappings as it was given in the original experiment's paper to create a suitable gamma-distribution. We controlled the number of participants for which data should be simulated and conservatively assumed that each participant produces only 100 valid trials (instead of the 120 trials actually performed in the experiment). For a fixed number of participants, we then repeated the following steps 60 times:

- set new random seed
- simulate new dataset given the number of participants
- update the pre-compiled model with the new data
- test hypothesis 1 based on the model
- check whether the Bayes factor is above the predefined threshold of 25
- calculate the new ratio of times we accepted the hypothesis

We obtained the following results:
For 2 participants the power of our model is 0.7
For 3 participants the power of our model is 0.867
For 4 participants the power of our model is 0.96667

Conclusion: A power of 0.8 is commonly assumed to be sufficiently large. Using this threshold, we would conclude that 3 participants provide enough power. This result was at first rather surprising for us since 3 seems to be a very small number of participants. We suppose the large number of trials per participant results in a sufficiently large dataset to detect a significant difference in the simulated reaction times of both mappings. But this does not mean that we plan to stop to collect data after the third participant. Other hypotheses might require a much larger amount of samples and especially the hypotheses regarding left-handed participants are probably harder to assess since the proportion of left-handed individuals is much smaller in the general population. We will therefore still try to recruit as many participants as possible but definitely at least 3. For details see the power_analysis script in our github repository.

15. **Stopping rule**
    We will stop collecting data (after four days of data collection) on the 3rd of August 2020 at 11:59pm.

Variables

16. **Manipulated variables**
    We manipulate the stimulus-response mapping: Either the response for a small stimulus is a left key ("q") and the response for a large stimulus is a right key ("p") or vice versa. We label the former (left=small, right=large) as the "compatible" and the latter (left=large, right =small) as the "incompatible" mapping condition.

17. **Measured variables**
    The response of the participant to a given stimulus will be recorded (either left "q" key or right "p" key) as well as the reaction time it took them from being shown the stimulus to them pressing the key (in ms). Additionally, we ask for the participant's handedness.

18. **Indices**
    We are not going to use any indices.

Analysis Plan

19. **Statistical models**
    We will analyze our data with regard to the seven hypotheses we defined previously. To assess hypotheses 1-6, we defined two identical Bayesian regression models. They analyze the influence of the factors S-R Mapping and Response separately on the (log-) Reaction Time as well as the interaction effect of S-R Mapping and Response on (log-) Reaction Time. We additionally accounted for group level variation between participants by adding random effects, namely random intercepts, and slopes for the factor Response.
    We will use the R package 'brms'. Our first two models (model_right and model_left) are defined by the formula:
    > logRT ~ Mapping * Response + (Response || Participant)

    For both models, the data is assumed to be Student-t-distributed and we did not replace the flat priors. Every other parameter of the models remains on the default setting. The models will run on the preprocessed data as described in section 22. data exclusion, however model_right will only be fitted with data from right-handed participants, while model_left will only use data from left-handed participants for fitting. The exact analysis can be found in the attached script.

    To assess hypothesis 7, we defined an additional model (model_both) which only slightly differs from the two previous models and is defined by the formula:
    > logRT ~ Mapping * Handedness + (1 || Participant)

    The handedness of the participants replaces the response location as a factor influencing (log-)Reaction time. Random effects now only comprise a random intercept. This model will use as dataset containing the exact same number of left- and righthanded participants. The family is again set to Student-t and the priors are left on the default setting (flat priors).
    The independent, manipulated variable is the S-R Mapping with levels compatible and incompatible, the further independent variable Response with levels left and right. These factors were encoded using the default dummy/treatment coding with the group level combination "compatible" and "left" as the reference level. For Model 3, a further independent variable is the Handedness, dummy coded with reference level "righthanded". The dependent variable is the (log-)Reaction Time.

20. **Transformations**
    We will log-transform the Reaction Time: logRT = log(RT).

We need to do this because Reaction Time measurements are typically assumed to be log-normally distributed and our regression model expects normally distributed data.

21. **Inference criteria**
    We will use the 'hypothesis' function from the 'brms' package and test our hypotheses with the posterior samples provided by our previously defined Bayesian regression models. This yields us a Bayes factor for each hypothesis (provided by the Evidence Ratio of the function output), which we compare to the interpretation table as suggested by Harold Jeffreys. Since this is a replication study, we picked a high threshold of 25. If the Bayes factor lies below this threshold, we will reject the respective hypothesis, otherwise we will fail to reject it. For more details on how the hypothesis function was employed, see the analysis script.

22. **Data exclusion**
    First, we will only use data from the experimental trials, i.e. exclude the data collected during training and distractor trials. Second, we only use trials in which participants responded correctly to the stimulus. Third, we will exclude Reaction Times faster than 100ms or slower than 1500ms.

23. **Missing data**
    If less than 10 trials were recorded incorrectly for a participant, e.g due to technical issues, we will use the remaining intact data from that participant. Otherwise we will discard all data from that participant.

24. **Exploratory analysis**
    If hypotheses (1)-(7) do not lead to significant results, we might look at credible intervals to determine if the effect occurs in opposite direction.

Other

25. **Other**
    Link to our github repository for the described study:
    https://github.com/Group1XPLab2020/Size_and_Space_exp1_magpie_Replication

26. **References**
    Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*, 371–396. DOI: https://doi.org/10.1037/0096-3445.122.3.371

    Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 626–641. DOI: https://doi.org/10.1037/0096-1523.16.3.626

Henik, A., Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Mem Cogn 10*, 389–395. DOI: https://doi.org/10.3758/BF03202431

Ren, P., Nicholls, M. R., Ma, Y., & Chen, L. (2011). Size matters: Non-numerical magnitude affects the spatial coding of response. *Plos ONE, 6*. DOI: https://doi.org/10.1371/journal.pone.0023553

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends In Cognitive Sciences, 7*, 483–488. DOI: https://doi.org/10.1016/j.tics.2003.09.002

Wühr, P., and Segeelke, C. 2018 Compatibility between Physical Stimulus Size and Left-right Responses: Small is Left and Large is Right. *Journal of Cognition, 1(1)*: 17, 1–11. DOI: https://doi.org/10.5334/joc.19