PCI Registered Reports #794

Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment

Stage-1 Submission, Version 3

Dear Mario,

Thank you for handling our submission and for letting us resubmit a revised version of our registered report "Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment".

In the following, you can find our replies to the comments. We revised the manuscript accordingly and highlighted all related changes.

While revising the manuscript, we noticed that we had never submitted the detailed literature overview on MC that we had prepared, although having referred to it in our manuscript (page 14): "see Table A1 in Appendix A". We apologize for bringing it up only now. We have appended it to the revised manuscript (pages 55 to 69). The table provides a thorough overview about studies that have investigated MC in the past, including information about their samples, used stimuli, continuous or categorical predictors. It also includes reported estimates for the correlation between the continuous MC-SNARC and PJ-SNARC for those studies in which PJ was assessed as well.

Best wishes,
Lilly (on behalf of all authors)

I have received the comments from the three reviewers and would like to thank them for their time and effort.

Two reviewers are pleased with the revisions you made. One noted a minor issue with grammar and bracket positions, which you may want to address. The other had no further comments and recommended acceptance.

The remaining reviewer suggested considering a potential issue with variability in category boundaries in the parity-judgment task. He proposed two approaches to explore this but left the decision to you. As the editor, I also leave it to you to address this point as you see fit. For the next revision, I will seek further input from this reviewer to assist in reaching a final decision regarding this first stage.

I look forward to your next version.
by Mario Dalmaso, 08 Jan 2025 16:24

We are glad that all three reviewers are satisfied with our previous revision. We find the third reviewer's (Peter Wühr's) new suggestions very valuable. As you will see in our

response to him further down, we have carefully considered his thoughts regarding the potential variability of boundaries between the categories of “small” and “large” numbers.

Note that we also made additional slight changes in the list of exploratory analyses (pages 19 and 20 in the track-change version). Apart from the newly added exploratory analyses about varying categorical boundaries and retesting the shape of the SNARC effect taking these into account (Exploratory 4, as proposed by Peter Wühr) we have only structured the list better and not changed anything content-wise.

Review by Michele Vicovaro, 02 Jan 2025 10:32

I thank the authors for having carefully considered and successfully addressed all of my previous comments. I just have a very minor point: there is a need to check the grammar and the bracket positions in the last sentence before the "SNARC and MARC compatibility" section on page 7 (in the manuscript with tracked changes).

Thank you for the opportunity to review this very interesting work. I look forward to seeing the full results!

We thank the reviewer for carefully checking our revision and for spotting the erroneous sentence, which we have corrected. We are glad that the reviewer found our work interesting, and we thank him for reviewing it.

Review by Christian Seegelke, 18 Dec 2024 19:09

The authors have sufficiently addressed all my comments and I happily recommend acceptance of this proposal.

We are glad that we sufficiently addressed the reviewer’s thoughtful comments, and we thank him for recommending our registered report.

Review by Peter Wühr, 15 Dec 2024 11:58

Summary: This is the revised version of a proposal for a highly powered online experiment addressing similarities and differences of SNARC effects in different tasks. The major aim of the study is to investigate possible differences in the effect functions of SNARC effects in two tasks, the magnitude-classification task (MCT) and the parity-judgment task (PJT), in a highly powered online experiment. In particular, the authors seek to investigate whether the effect functions, which describe the relationship between numerical magnitude (of stimuli) and RT differences between left and right responses, are categorical or continuous in nature.

Evaluation: I already liked the previous version of the proposal. Moreover, in their revision, the authors have adequately addressed my comments on the previous version of the proposal, and incorporated many comments and suggestions made by other reviewers. As a result, in my view, the quality of the proposal has improved over the previous version.

> We are glad that we have adequately addressed the comments and that our registered report has improved in the reviewer's view.

However, after having submitted my previous review, a new account for the apparently continuous shape of the effect function in the PJT came to my mind, and I would like to share this idea with the authors. Being an author myself, I know that it can be frustrating for authors to receive reviewer suggestions in such a piecemeal manner, but I think this additional comment may be worth a thought before starting data collection.

Previous research suggests that the effect functions, which describe the relationship between numerical stimulus magnitude and RT differences between left versus right responses, may have different shapes for MCT and PJT. In particular, the effect function seems to be categorical (step-like) in MCT, with the boundary separating small from large numbers at 5. In contrast, the effect function appears to be continuous (linear) in PJT. The authors wish to test between the two possibilities by comparing the fits (R^2) of a continuous and a categorical model to the individual data obtained in both tasks.

A critical assumption of this approach is that, when the "true" effect function is categorial, the boundary separating the small stimuli from the large stimuli is the same for all participants in both tasks, that is, the median stimulus value 5 is always the boundary. This assumption, however, is not necessarily true. Rather, it is possible that the effect function in both tasks—that is, in MCT and in PJT—is categorical, with the boundary being the same for all participants in MCT, whereas the boundary varies across participants in PJT. In my view, it makes sense to assume that the boundary is more consistent across participants in MCT than in PJT because the boundary is explicitly defined or instructed in MCT, but not in PJT. Hence, it is possible that, in PJT, the boundary that separates the small from the large numbers is smaller than 5 for some participants, and larger than 5 for others. When averaging across individual data, which reflect categorical functions with variable boundaries, the misleading impression of a gradual effect function would arise. A similar problem has been identified and discussed in the literature on learning curves (e.g., Bahrick et al., 1957; Gallistel et al. 2004).

> We sincerely thank the reviewer for sharing these thoughts, which we find very insightful. Boundaries for categorizing numbers into "small" and "large" might indeed vary between individuals, although we agree that this is more likely in PJ than in MC, because 5 is explicitly defined as a boundary in MC task instructions. We also fully agree that aggregating data over participants can blur individual differences and may result in an inaccurate overall picture.
>
> Towards the end of the subsection "Different shapes of the SNARC effect" within the Introduction (page 15), we have therefore inserted the following new paragraph:

"*Importantly, the boundary between "small" and "large" numbers is not necessarily 5 for every individual and might instead vary between individuals. A split of the full number interval into two halves (i.e., from 1 to 4 and from 6 to 9) seems plausible in MC, where the boundary of 5 is explicitly defined in the task instructions. However, especially in PJ, but potentially even in MC, some individuals might classify numbers into the categories "small" and "large" with a different boundary. Crucially, an overall continuous SNARC effect could result either from continuous patterns in most individuals, or from averaging across categorical patterns differing between individuals. In this study, we will therefore additionally determine the most likely categorical boundary for each participant separately. Subsequently, we will investigate the shape of the SNARC effect (Hypotheses 1 and 2) once more, this time comparing the best fitting categorical model for each individual with the continuous model (Exploratory 4).*"

Moreover, we have added the new suggestion to the list of explorations (pages 19 and 20) and refer to it as Exploratory 4 throughout the manuscript. Specifically, we wrote that we want to explore whether the following observation can be made:
"*number 5 as the most likely boundary between "small" and "large" numbers for most individuals both in MC and PJ; and, relatedly, a categorical MC SNARC (Hypothesis 1) and a continuous PJ-SNARC (Hypothesis 2), even when comparing the most likely categorical model for each participant with the continuous model*"

I am not sure if the methodological approach of the authors for testing between continuous versus categorical effect functions, which assumes a constant boundary between categories at 5, is appropriate (or sensitive enough) for detecting categorical functions with variable boundaries. This issue might be worth a thought. In the following, I will briefly sketch two ideas for testing for categorical functions with variable boundaries. The authors may have better ideas on this issue, or they may find better methods in the literature.

The analysis we originally proposed did not account for the possibility of variable boundaries between "small" and "large" number categories. To our knowledge, previous studies have not investigated this possibility either. We agree with you that this issue is worth including in our manuscript!

A first possibility would be to use a larger range of categorical (step) functions, with variable boundaries between 3-7, that are fitted to the individual data sets. A problem with this approach might be that, since there are more categorical than linear models, categorical models might have a higher chance for gaining a better fit than the linear model.

A second possibility would be to analyze and compare values and slopes at the extreme ends of the stimulus set. This approach rests on the assumption that variable boundaries are very unlikely to occur at the extreme ends of the stimulus range (i.e. between 1 and 2, or between 8 and 9). If this assumption holds, a categorical function would predict that slopes between 1 and 2 and between 8 and 9 were both zero, while the criterion variable differed between both ends of the stimulus range. In contrast, a continuous function would predict that the slopes between 1 and 2 and between 8 and 9 were both different from zero, and of similar size.

We thank the reviewer for suggesting two approaches of how to test variable boundaries and compare the most likely categorical model with the continuous one. Of the possibilities proposed by the Reviewer, we favored the first one, which we decided to apply after some modifications, including additional check by means of split-half reliability to figure out whether what we capture represents a consistent pattern within individuals. Since we find this idea not only interesting but also plausible, we added an exploratory analysis (page 34 to 36):

"*Then, we will fit five categorical models for each participant per task. The models will be analogous to MC-2 and PJ-2. They will differ regarding the boundary between "small" and "large" numbers (contrast-coded with -0.5 and +0.5, respectively). Specifically, we will run five regression models while classifying 1 to 2 vs. 3 to 9, 1 to 3 vs. 4 to 9, 1 to 4 vs. 6 to 9, 1 to 6 vs. 7 to 9, and 1 to 7 vs. 8 to 9 as "small" and "large" numbers (see Table 1). Note that 1 to 4 vs. 6 to 9 corresponds to the model used to test Hypotheses 1 and 2. For each participant, we will determine the most likely underlying categorization by descriptively comparing which of the five models has the best fit to the data in terms of* R^2*. Next, we will logit-transform the* R^2 *for the favored categorical model and for the continuous model for each participant separately and compare the logit-transformed* R^2 *between the two models in a two-sided Bayesian paired t test (Exploratory 4). Moreover, we will regress dRTs on continuous and categorical magnitude for both PJ and MC in four separate Bayesian models and perform a leave-one-out cross validation (as for Hypotheses 1 and 2). We will present the distribution of favored categorical models across the sample for each task.*

Additionally, to check whether the determined boundary between "small" and "large" numbers is reliable for each participant and not only due to random measurement noise, we will determine its split-half reliability by splitting valid trials with the odd-even method based on presentation order (i.e., 1st, 3rd, 5th, etc. trial vs. 2nd, 4th, 6th, etc. trial). Subsequently, we will again compute five models per participant and per task, separately for each half of all valid trials, and determine which of the five models has the largest R^2 *in each half of the experiment. For each half per participant within each task, we will code the boundary of the model with the highest* R^2 *with the cardinal boundary values 2.5, 3.5, 5, 6.5, or 7.5 We will then compute the Pearson product-moment correlation between the favored categorical models across halves within each task (note that coding the favorable model ordinally with 1, 2, 3, 4, and 5 to calculate the Spearman rank correlation would lead to the almost same results, as the correlation between the cardinal and ordinal values is* $r = .997$*). Next, we will apply the Spearman-Brown correction to each correlation to adjust for task length and descriptively evaluate the correlations in terms of whether the boundaries between "small" and "large" numbers are related between both halves. As the boundaries can be also interpreted as categorical rather than continuous values, we will also create alluvial plots (using the R packages ggplot2 by Wickham et al., 2024, and ggalluvial by Brunson & Read, 2023) showing the stability of the boundaries within participants across halves for each task separately.*

Finally, we will compute the split-half reliability for the basic categorical MC-SNARC (using the classification of 1 to 4 vs. 6 to 9 as "small" vs. "large") and continuous PJ-

SNARC. These results will be taken into account when interpreting the split-half reliability of the favored categorical model (as a part of Exploratory 4). Note that, to our knowledge, the split-half reliability has never been reported for the categorical MC-SNARC. For the split-half reliability of the continuous PJ-SNARC, values between .43 and .82 have been reported in the literature (for an overview, see Cipora, Soltanlou, et al., 2019)."

We added that, if not specified otherwise, throughout our registered report, by "the categorical predictor" we mean the one that categorizes numbers from 1 to 4 as small and from 6 to 9 as large (page 14; see also Note to Table 1). In some places, we also changed "**the** categorical model/predictor" to "**a** categorical model/predictor", because we want to make clear that there are more possible boundaries than only 5 (e.g., page 14, page 19). Moreover, when referring to previous studies yielding evidence for a categorical shape in MC, we specify that this result comes from data that was averaged across participants (e.g., page 13).

Note that we acknowledge the reviewer's insightful feedback in the manuscript in Footnote 3 about boundaries between "small" and "large" potentially differing between participants (page 15) as well as in Footnote 4 about predictions derived from the dual-route model (page 16). Please feel free to let us know in case you do not wish to be mentioned as a stage-1 PCI-RR reviewer in the manuscript.