PCI Registered Report

SNARC flexibility (e-SNARC study 3)

Dear Mr. McIntosh,

Thank you for giving us the opportunity to submit a revised version of our registered report manuscript with the title "How Flexible are Spatial-Numerical Associations? A Registered Report on the SNARC's Range Dependency". We are grateful for your and the two reviewers' valuable feedback and appreciate the helpful comments.

We must inform you that Gáspár Lukács is no longer working together with us on this project and does not wish to be included as a coauthor. We therefore deleted his name from the manuscript's front page and the author contributions. Further, we decided to change the title of our manuscript to "One and only SNARC? The Flexibility of Spatial-Numerical Associations. A Registered Report on the SNARC's Range Dependency"

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

Kind regards,

Lilly Roth (on behalf of all authors)

Thank you for your patience, and I apologise that it has taken so long to return reviews for your manuscript. It proved tricky to find reviewers, but once suitable and willing people were found, the reviews were thorough, and I think you will find them very helpful in guiding a revision of this Stage 1 plan. Reviewer#2 in particular is very familiar with the logic and rigours of Registered Reports, and provides excellent guidance on related considerations.

Thank you for your effort to find suitable and willing reviewers for our manuscript.

Both reviewers are generally positive about the proposed study, but have substantive concerns and suggestions for improvement. You should consider (and respond to) all of these points carefully. I would emphasise the following in particular, adding some comments of my own.

Reviewer#1 (Melina Mende) makes a number of requests for clarity, and emphasises the need for a full justification for the approach to data trimming. The approach to treatment of Reaction Times is absolutely critical (since it is the basis for the core dependent measure). It is described in detail, but it is not rationalised. The treatment of RT is a complex issue, and decisions about whether (and how) to exclude outliers and/or to transform data and/or to use robust estimators of central tendency per cell (i.e. medians) are complex and ideally should be informed by a good working knowledge of the characteristics of the data in your experiment. (In general, of course, pre-registered approaches may tend towards more conservative and robust approaches, because the final form of the data cannot be known in advance.)

Thank you for emphasizing the need for justification of our data preprocessing as raised by Reviewer 1. We have added explanations for the respective steps to the

revised manuscript in the paragraph *Data preprocessing* and believe that the rationale behind these steps is now better comprehensible for the readers. We have also responded to Reviewer 1 in comment 10 with the following points:

a) *Exclusion of RTs below 200 ms: We have added to the manuscript that "parity judgments faster than 200 ms are very unlikely and faster responses can therefore be treated as anticipations". We do agree that a threshold of 150 ms or 250 ms would be equally justified as the one we chose. Our experience from reanalyzing several SNARC studies from different labs shows that well below 1% of the data (e.g., Cipora, van Dick et al., 2019) is excluded from the analysis with a threshold of 200 ms.*
b) *Exclusion of RTs above 1500 ms: We do not plan to implement a trial timeout in our online study (which is quite typical in lab studies; see Table 2 in Cipora et al., 2019, for examples of some task parameters). The reason is that the participants could be distracted and stressed after missing a trial. Therefore, the dataset will contain some extremely long RTs, which do not reflect the actual mental process underlying parity judgment. Therefore, before any further trimming, we exclude these RTs. According to our experience 1500 ms seems to be a reasonable border to consider in case of parity judgment of single digit numbers, and as an explanation, we have added to the manuscript that "healthy educated adults should be capable to judge the parity status of single-digit numbers in less than 1500 ms, so that slower responses are unlikely to reflect only the mental process underlying parity judgment but instead might be caused by distractions". As above, we acknowledge that 1500 ms is arbitrary, and one may argue whether any other value should be used instead (e.g., Gevers et al., 2006 used 800 ms as a threshold). Nevertheless, a decision needs to be taken and we opt for this. We used that same criterion earlier in Cipora et al., (2019), where we received SNARC effects of similar sizes as in lab studies. Additionally, the same trimming criteria allow for direct comparisons with that earlier online experiment if this deems necessary in some post-hoc exploratory analysis for any unforeseen reason. Note that this general fixed threshold can be rather liberal, because it is complemented by individualized sequential trimming (see below), which eliminates RTs that are not likely to be part of the underlying distribution of interest.*
c) *Sequential trimming within participants: We have done this in multiple earlier studies (since Nuerk et al., 2001) and find the technique reasonable. Namely, as we have added to the manuscript, "this procedure permits to exclude RTs that are unlikely for each given participant and accounts for the right-skewed distribution of RTs, where the means would otherwise be largely overestimated". By calculating the mean and standard deviations for the sample of trials, we wish to estimate the underlying true mean and standard deviation. However, these estimators are systematically biased if trials are included that are not part of the true underlying distribution, such as slower responses due to distractions of the participants (the underlying distribution of such trials would be not only the distribution of parity judgements, but also the unknown distribution of time delays due to distractions). Therefore, means and standard deviations are recalculated after eliminating trials that are highly likely not part of the underlying true distribution.*

d) *Our attempt at multiverse analysis of SNARC data (work still in progress) has shown that the effect of specific data preprocessing routine on finding the SNARC is relatively small unless an attempt is being made to exclude very large proportion of trials (e.g., sequential trimming with ±2SD). Note that we use ±3SD, so that 99.9% of the values should lie within the trimming borders assuming a true normal distribution underlying the data.*

e) *Excluding participants with fewer than 75% valid trials: Parity judgment with single-digit numbers is very simple. It is a forced-choice reaction task with a chance of 50% for a correct response. We only want to analyze data from participants who most likely completed the task following the instructions and believe that a criterion of half of the trials above chance level is reasonable. Therefore, we selected a criterion of 75% valid trials, which implies on average 50% correct responses and 50% blind guessing (thereof 50% erroneous and 50% correct responses, as well as some very fast or very slow responses). Again, as above, we acknowledge it is an arbitrary decision, and a threshold of 70% or 80% would also be adequate.*

f) *Excluding participants with empty cells (number * response side): As we explain in the revised manuscript, "an empty cell causes a missing dRT, which in turn makes the calculation of the SNARC slope problematic".*

On this point, although this Stage 1 RR seems to have been thought through carefully, I do not see direct evidence of the tasks having been fully piloted, where pilot data would allow for the full piloting of the proposed analysis pipeline. It may be that you have done such piloting, or perhaps you have used the same data collection approach/platform in a previous study, so your analysis pipeline is well established. If so, then you should describe this relevant history in the present RR. If not, then I honestly think it is necessary to conduct a reasonably-sized pilot in order to debug and optimize your analysis plan.

We have added that "The parity judgment task is widely used in numerical cognition and the standard task to investigate the SNARC effect (see Toomarian & Hubbard, 2018, for a review, and Wood et al., 2008, for a meta-analysis)." to the paragraph *Design and experimental task*. Importantly, the parity judgment task has been used in the studies we aim at replicating. The chosen length of the tasks was guided by our previous experience with SNARC measurement. More precisely, we will use 25 repetitions per experimental cell and cite the simulation study by Cipora and Wood (2017) stating that at least 20 repetitions per experimental cell are needed in the paragraph *Methodological limitations of the two initial studies demonstrating RMdependency*. Moreover, we have added to the paragraph *Data preprocessing* that our proposed preprocessing pipeline "is similar to that used by Cipora, van Dijck, et al. (2019) in an extensive re-analysis of 18 datasets and permits to reliably detect the SNARC effect." A very similar pipeline has also been proven to be suitable for detecting the SNARC effect in the largest SNARC study so far in an online setup (Cipora, Soltanlou, et al., 2019). Finally, we have added to the paragraph *Procedure* that our previous studies have demonstrated that the software is suitable to detect the SNARC effect online.

To sum up, we therefore do not believe that we need yet another pilot study. However, we agree that we did not have enough justifications and background when proposing the experimental task, the data collection approach, and the analysis pipeline in the previous version of our RR. Therefore, we acknowledge that the suggestion to run

another pilot was reasonable based on this previous version. However, we are convinced that because of all the pilot work already conducted by us and by others as outlined above, there is no need of more piloting. We hope that this is now more acceptable for the readers, the reviewers and the recommender.

This relates also to the points made by Reviewer#2 regarding your quality checks (and confidence in the is capable of testing the hypotheses of interest seem essential here.

To our revised RR we added a new subsection with the header *Positive controls and manipulation checks* (as proposed by Reviewer 2, comment 5):

"To control the data quality in our study, we have implemented a seriousness check (Aust et al., Reips, 2009, review in Reips, 2021) as well as a self-assessment of noise, distractions, and other difficulties. To make sure that we will only analyze trials that reflect mental processes in correctly executed parity judgment, we will exclude incorrectly answered trials and trim RTs (as described in the data preprocessing pipeline). Also, we will exclude full datasets of participants with less than 75% valid trials to only build our results on participants who have understood and followed the task instructions. Moreover, we ask participants whether they complied with the instructions to use their left and right index fingers for the left and right response keys, respectively.

Last, we will check for the presence of the Odd Effect (Hines, 1990; i.e., overall faster reactions to even than to odd numbers, irrespective of the response side). The Odd Effect is quite robust in the parity judgment task, but independent from the SNARC effect (as it is independent from number magnitude and from its mapping onto space and only considers parity). Therefore, we can consider its investigation as a manipulation check, and in case of its presence we will have a positive control for our experiment. For this, we will subtract the average RT for even numbers from the average RT for odd numbers per participant and test the differences (one per participant) against zero in a Bayesian one-sample *t*-test (with positive estimates indicating the Odd Effect)."

I also agree strongly with this reviewer that the purpose and inferential role of all parts of the analysis must be clear (e.g. how will the outcome of the dropout analysis be used to inform interpretation of findings), and that the exploratory analyses should be removed from the Stage 1 plan. The 'follow-up' analyses should be elevated to full inferential status and specified fully or, if not essential to the main conclusions, relegated to exploratory status and omitted (the latter approach may be preferred as simpler, given that your analysis plan is already rather complex).

We have removed the entire paragraphs *Follow-up analysis* and *Exploratory data analysis* from the manuscript.

I also concur with the idea that combining frequentist and Bayesian approaches seems unnecessarily complex and ambiguous. If these approaches do not lead to the same outcomes then which approach will you be guided by? (And then why should you bother to include the other approach at all?) It is of course possible to include parallel frequentist and Bayesian analyses in an RR, but specifying unambiguously which theoretical conclusions will follow for the full range of possible outcomes becomes very complex.

Thanks for this critical and valuable feedback. We have decided to remove the frequentist analyses and instead go for the "Sequential Bayes factor with maximal n" approach (Schönbrodt & Wagenmakers, 2018), as we explain in our reply to Reviewer 2 in comment 9:

Thanks for this helpful comment. We have decided to employ only the Bayesian approach instead of combining it with the frequentist approach. We appreciate your suggestion to use sequential analyses with optional stopping and decided to go for the "Sequential Bayes factor with maximal n" (SBF+maxN) approach as suggested by Schönbrodt & Wagenmakers (2018).

We have revised the section Statistical power considerations and sample size determination in our manuscript and explained the SBF+maxN approach there. Also, we have revised the Study Design Table. Namely, we have set the minimum sample size to 200 participants and will analyze the data after each 20 new datasets until we reach moderate evidence against or for AMdependency of the strength of the SNARC effect (i.e., threshold of $BF_{10} > 3$ or $BF_{10} < 1/3$ for or against Hypothesis 3). We have chosen to make the optional stopping dependent on Hypothesis 3 because it is the most crucial hypothesis in our study. Moreover, a sample that is large enough to find moderate evidence against or for Hypothesis 3 is most probably also large enough to find at least moderate evidence for the SNARC effect in different ranges (Hypothesis 1) and for both RMdependency and AMdependency of the number mapping on the MNL (Hypothesis 2). We have defined a maximum sample size of 700 participants, which is the sample size that we have determined to be necessary for detecting an effect of Cohen's d = 0.15) with power of .90 when using the rather large standard deviations observed by Fias et al. (1996). Note that although the term "moderate evidence" does not sound very convincing, the respective BF_{10} threshold of 3 corresponds to p-values below .01, so that we consider this to be a rather conservative and adequate threshold.

With regard to the frequentist analysis, I have some concerns about the approach to apha levels and (non-)adjustment for multiple comparisons. In the text you state: "*For each test described below, a significance level of $\alpha = .01$ will be used. The reason for using a rather conservative significance level is that we will conduct multiple tests per hypothesis… Importantly, the significance level does not need to be corrected for the total number of conducted tests in this study, because the tests belong to different test families and because different theoretical inferences can be drawn from their results (Lakens, 2016). Moreover, we will look at each result individually and not generalize from one single significant result within a test family to the presence of an effect in both experiments and in all possible number ranges, so that our interpretations will not inflate the familywise error rate.*"

This sounds very thorough, but I am not sure it is sound/coherent. First you state that you adopt a conservative significance level so that you don't have to adjust for multiple comparisons – it would be more transparent to state what the significance criterion is, and how it has been adjusted for (how many) comparisons. Without this, it is unclear what your effective significance threshold is. In apparent contradiction to the above you then go on to state that the threshold does not need to be adjusted because the individual tests are all testing independent hypotheses, and you will interpret each individually. This logic is repeated in the design table.

Yes, you are right, our proposal was not coherent. Thank you for pointing this out. In our revised Registered Report, we run only Bayesian analyses, so we do not need to

define significance levels. Instead, we will report whether we find moderate ($BF_{10} < 1/3$ or $BF_{10} > 3$) or strong ($BF_{10} < 1/10$ or $BF_{10} > 10$) evidence.

Although this approach may seem appealing, I am not sure that it is convincing in the present case. As far as I am aware, there is no theory proposing that functionally independent SNARC effects may exist for your different number ranges.

We do not know of any theory proposing that SNARC effects entirely depend on the number range (AMdependency), but as outlined in the introduction of our Registered Report, we do not believe that SNARC effects are entirely flexible either (RMdependency). As further argued in the introduction, there is, however, a general tendency to interpret the SNARC effect as entirely flexible (RMdependency) based on the findings of range dependency of dRTs and on the inference-statistical null effects of SNARC slope comparisons between ranges in the original underpowered studies. Hence, suggesting AMdependency of the SNARC effect based on these seminal studies is actually a different view from dominant accounts.

Finally, one of us (Nuerk) has proposed that the SNARC effect operates on multiple number ranges in the paper „Attention allows the SNARC effect to operate on multiple number lines" (Weis et al., 2018). However, that paper is not about whether the SNARC effect operates on multiple number lines in terms of RMdependency and AMdependency, but it used two-digit numbers as stimuli to see whether separate number lines are activated for decade and unit numbers. There, the operations on different number ranges are for decade and unit digits of one two-digit number (i.e., the same number, but different digits of its decomposition). The current RR goes far beyond that because it seeks to demonstrate that both RMdependent and AMdependent spatial mappings are present concurrently in the same digit. We have added that point to our revised manuscript and think this makes the theoretical point more concise.

In any case, you also state that "*not finding it [the SNARC effect] in one of the four ranges despite our large sample would speak against the robustness of the SNARC effect*". This means that the results are not really being evaluated independently for each number range, but considered together to bear on the same theoretical question. Moreover, it is not convincing to state that the failure to find the result in one of the four ranges would speak against the robustness of the SNARC effect, because 90% power implies a 10% chance of a false negative in any one range (~40% chance of at least one false negative result).

Thank you for pointing this out, this is true, the failure of finding the SNARC effect in one of the ranges would not speak against its robustness. We have now changed all frequentist analyses to fully Bayesian analyses, which can help us differentiate evidence of absence from absence of evidence here. We have replaced the sentence "not finding it [the SNARC effect] in one of the four ranges despite our large sample would speak against the robustness of the SNARC effect" with "We […] expect to find at least moderate evidence for it [the SNARC effect] in all four number ranges. Finding at least moderate evidence against the SNARC in any of the four number ranges would be highly surprising given that it is a robust effect in the parity judgment task."

In general, I think that your statistical approach needs better justification and specification, and that it might benefit from simplification (e.g. by deciding on either a frequentist or Bayesian approach). In passing, I note that you refer to another paper (Roth, Lukács, et al.,

2022) for your power calculations (which, confusingly, seems to be an earlier version of this same RR plan). In any case, power calculations are an integral part of a Stage 1 RR plan and so should be described fully in the RR itself.

> The cited power calculations are not part of another paper or of an earlier version of this same RR plan. Instead, the power calculations we refer to are the ones we run for this current RR and posted on OSF as a supplementary material to this submission as a PDF-rendered R Markdown file. We are sorry for the misunderstanding. Instead of citing it and listing it in *References*, we have now replaced the citation by the DOI URL when referring to our calculations in the manuscript. Similarly, we have deleted the citations and the *References* entry of our preregistration for the color judgment studies and only put the DOI URL into the manuscript text.

I hope that the reviewers' helpful comments will be useful to you in taking this project forward, and if you decide to revise this Stage 1 RR, then you should indicate how you have responded to each of the comments made, including the additional ones above.

> Yes, your comments and the comments written by the two reviewers were definitely helpful. Thank you for the valuable and constructive feedback.

Review 1 by Melinda Mende

The article is well-written and targets an interesting and relevant issue. The researchers aim to investigate RMdependency and AMdependency of the SNARC effect. I think that this work is a positive example of a well-designed study where a lot of considerations were made, starting with the optimal number of trials per cell and power calculations. Further, also the planned data analysis is well described with useful measurements to improve the quality of the statistical analysis. Overall, I have just some minor suggestions to further improve this work.

> Thank you. We are happy to hear that you find our study interesting, relevant, and well-designed.

1. p.7: “In that study, the observed result pattern looked like Scenario 5 in Figure 1”.

   Figure 1 is too far away from this claim. I suggest either introducing the figure earlier or not referring to it at this stage.

   > In the revised version of our RR, we do not refer to Figure 1 at this stage yet. In any case, when we describe the scenarios later in the text (pages 10 and 11), we also say what Dehaene et al’s (1993) and Fias et al.’s (1996) results looked like, so we can skip it at this stage.

2. p.7: The content of footnote 1 would be better in the main text together with the previous explanation on how to calculate the SNRAC effect.

   > We have moved the content of footnote 1 to the main text.

3. p. 10/11: Scenarios are very hard to understand, even though they are illustrated in Figure 1. One needs to scroll up and down a lot. Maybe you could divide the figure into parts and explain each of the scenarios and then directly show the figure.

We have divided the original figure that included all six scenarios into two figures. In the revised Registered Report, Figure 1 contains Scenarios 1, 2, and 3, and directly follows the description of these scenarios in the text, and Figure 2 contains Scenarios 4, 5, and 6, and directly follows the description of these scenarios in the text. This way, readers do not have to scroll up and down so much. However, we would like to keep Scenarios 1, 2, and 3 together, because they all display AMdependency and RMdependency of the number mapping on the MNL, and we would like to keep Scenarios 4, 5, and 6 together, because they all display AMdependency and RMdependency of the strength of the SNARC effect. Displaying them next to each other facilitates the direct comparison of scenarios.

4. p. 12: "namely 0 to 5 and 4 to 9 in Experiment 1, and 1 to 5 (excluding 3) and 4 to 8 (excluding 6) in Experiment 2"

   Which study are you referring to?

   We are referring to the number magnitudes in our own study that we plan in this Registered Report. Thanks for pointing out that this was unclear, we have now specified this in the text.

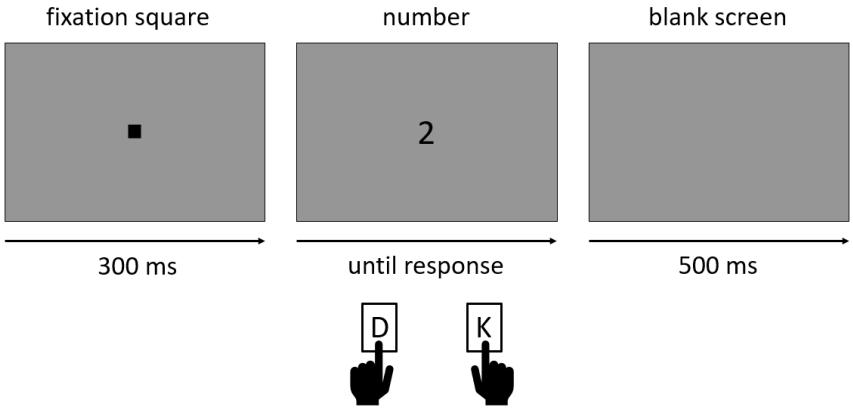
5. p. 16: I do understand your design approach and I think that the two experiments are well elaborated. Nonetheless, I do not understand the content of Table 2. What do you mean by "Parity +0.5"/"Parity -0.5"?

   As explained in the *Note* for Table 2, parity is contrast-coded with -0.5 for odd and +0.5 for even numbers. To avoid confusion, we have changed the column headers from "Parity" to "Contrast-coded parity predictor".

   The table illustrates that the dRT predictors number magnitude and number parity for the repeated-measures regression are not orthogonal to each other in Experiment 1, which is a replication of Dehaene et al.'s (1993) and Fias et al.'s (1996) experiments using their exact stimulus set, because the mean number magnitude is not independent from number parity and because the correlation between these two predictors is not zero. Therefore, in the conceptual replication that we conduct as Experiment 2, the two predictors are orthogonal to each other, so that their contribution can be truly interpreted as independent from one another.

6. p.17: Why not visualize the time course of the stimuli presentation with a Figure?

   Thanks for the suggestion. We have created a new Figure 4 to visualize the time course of the stimuli presentation:



7. p. 18: “This figure shows the four between-subjects conditions”

   Why don’t you want to use a fully within-subject design?

   Talking about “four between-subjects conditions” was wrong here, we are sorry for that. In fact, we do use a fully within-subject design. We talk about “four counterbalanced block orders” in the revised Registered Report to avoid misunderstandings. Thanks for pointing this out.

8. p. 18: “handedness, and finger-counting habits”

   How will these be measured?

   In the revised Registered Report, we explain how we measure handedness and finger-counting habits that we had not described in the first version. For handedness, the answer options will be *right-handed*, *left-handed*, *ambidextrous (two-handed)*, and *I prefer not to answer*. For finger counting habits, participants have to indicate with which hand they start counting by choosing *left*, *right*, *I do not know or I do not have any preferred hand*, or *I prefer not to answer*; and next they can indicate the stability of this preference by choosing *always*, *usually*, *I do not know or I do not have any preferred hand*, or *I prefer not to answer*.

   Note that we also provide links to demo versions of both experiments, such that readers can check out the exact questionnaire (Experiment 1 at https://luk.uni-konstanz.de/numcog_3/?demo&e1 and Experiment 2 at https://luk.uni-konstanz.de/numcog_3/?demo&e2).

9. p. 18: “Participants may choose response keys for the experimental task which are to be located in the same row and about one hand width apart from each other on their keyboard”

   Even if the keys are one hand width apart from each other, how do you make sure that participants do not use just one hand for giving their responses?

   We implemented the option to use other response keys than the default (D and K as left and right response keys) in case participants in this internationally conducted online experiment use an unconventional or foreign type of keyboard, where D and K are not in the same row or not approximately one hand width apart from each other. With this procedural improvement, we can avoid potential confounds (e.g., if the left

response key is above the right response key, there would be a confound between left-right directionality and up-down directionality).

We must admit that we cannot guarantee that participants use the index finger of their left hand for the left response key and the index finger of their right hand for the right response key. Controlling for this is difficult in an online study, no matter whether participants have the option to change the default response keys or not (as a side note, it is also not entirely feasible in lab studies, as this would require the experimenter to constantly look at participants' hands during the experiment, which in turn could lead to other unexpected effects like the Hawthorne effect, where awareness of being observed alters behavior, and additionally administration of the experiment would be more difficult).

However, we instruct participants to use both hands, and we do not see any reason for them not complying with these instructions. Note that in a previous study (Cipora et al., 2019), we also found SNARC effects of similar size and direction in online studies as in previous lab studies – therefore, data from online SNARC studies do not indicate that hands are used in an inappropriate way. Given the distance of the two response keys on the keyboard (namely, approximately one hand with), it should even be uncomfortable for participants to use just one hand for both response keys.

To sum up, we think that this should not pose a problem. Nevertheless, we will add a control question at the end of the experiments, where we ask participants whether they have used their left and right index fingers to be even more sure that the experiment was conducted in an appropriate way. Note that this control question is not implemented in the demo versions yet, but we will do this as soon as possible.

10. p. 19: "Only trials with RTs between 200 and 1500 ms will be included in the analysis. Further outliers will be removed in an iterative trimming procedure for each participant separately, such that only RTs that are maximum 3 SDs above or below the individual mean RT of all remaining trials will be considered. Finally, only datasets of participants with at least 75% valid remaining trials and without any empty experimental cell (number magnitude per response side) in both number ranges will be considered."

    Please specify how and why you selected these criteria. Such data trimming criteria are often similar in the literature but not entirely equal. Thus, I would like to learn about your justification for using these criteria.

    This is a fair point. We have added justifications for using these criteria to the paragraph *Data preprocessing* of the revised manuscript. We developed this routine for dealing with RT data from our first SNARC online study. The logic behind is as follows:

    a) Exclusion of RTs below 200 ms: We have added to the manuscript that "parity judgments faster than 200 ms are very unlikely and faster responses can therefore be treated as anticipations". We do agree that a threshold of 150 ms or 250 ms would be equally justified as the one we chose. Our experience reanalyzing several SNARC studies from different labs shows that well below 1% of the data (e.g., Cipora, van Dick et al., 2019) is excluded from the analysis with a threshold of 200 ms.

    g) Exclusion of RTs above 1500 ms: We do not plan to implement a trial timeout in our online study (which is quite typical in lab studies; see Table 2 in Cipora et al.,

2019, for examples of some task parameters). The reason is that the participants could be distracted and stressed after missing a trial. Therefore, the dataset will contain some extremely long RTs, which do not reflect the actual mental process underlying parity judgment. Therefore, before any further trimming, we exclude these RTs. According to our experience 1500 ms seems to be a reasonable border to consider in case of parity judgment of single digit numbers, and as an explanation, we have added to the manuscript that "healthy educated adults should be capable to judge the parity status of single-digit numbers in less than 1500 ms, so that slower responses are unlikely to reflect only the mental process underlying parity judgment but instead might be caused by distractions". As above, we acknowledge that 1500 ms is arbitrary, and one may argue whether any other value should be used instead (e.g., Gevers et al., 2006 used 800 ms as a threshold). Nevertheless, a decision needs to be taken and we opt for this. We used that same criterion earlier in Cipora et al., (2019), where we received SNARC effects of similar sizes as in lab studies. Additionally, the same trimming criteria allow for direct comparisons with that earlier online experiment if this deems necessary in some post-hoc exploratory analysis for any unforeseen reason. Note that this general fixed threshold can be rather liberal, because it is complemented by individualized sequential trimming (see below), which eliminates RTs that are not likely to be part of the underlying distribution of interest.

b) Sequential trimming within participants: We have done this in multiple earlier studies (since Nuerk et al., 2001) and find the technique reasonable. Namely, as we have added to the manuscript, "this procedure permits to exclude RTs that are unlikely for each given participant and accounts for the right-skewed distribution of RTs, where the means would otherwise be largely overestimated". By calculating the mean and standard deviations for the sample of trials, we wish to estimate the underlying true mean and standard deviation. However, these estimators are systematically biased if trials are included that are not part of the true underlying distribution, such as slower responses due to distractions of the participants (the underlying distribution of such trials would be not only the distribution of parity judgements, but also the unknown distribution of time delays due to distractions). Therefore, means and standard deviations are recalculated after eliminating trials that are highly likely not part of the underlying true distribution.

c) Our attempt at multiverse analysis of SNARC data (work still in progress) has shown that the effect of specific data preprocessing routine on finding the SNARC is relatively small unless an attempt is being made to exclude very large proportion of trials (e.g., sequential trimming with ±2SD). Note that we use ±3SD, so that 99.9% of the values should lie within the trimming borders assuming a true normal distribution underlying the data.

d) Excluding participants with fewer than 75% valid trials: Parity judgment with single-digit numbers is very simple. It is a forced-choice reaction task with a chance of 50% for a correct response. We only want to analyze data from participants who most likely completed the task following the instructions and believe that a criterion of half of the trials above chance level is reasonable. Therefore, we selected a criterion of 75% valid trials, which implies on average 50% correct responses and 50% blind guessing (thereof 50% erroneous and 50%

correct responses, as well as some very fast or very slow responses). Again, as above, we acknowledge it is an arbitrary decision, and a threshold of 70% or 80% would also be adequate.

e) Excluding participants with empty cells (number * response side): As we explain in the revised manuscript, "an empty cell causes a missing dRT, which in turn makes the calculation of the SNARC slope problematic".

Review 2 by anonymous reviewer

The authors of the present Stage 1 Registered Report aim to investigate flexibility of Spatial-Numerical associations by means of two experiments, one being a close replication and one a conceptual replication of previous studies in the field. I think the topic is highly timely, given the accumulating evidence on SNA and its implications. Overall, the authors have obviously taken great care in reviewing the existing literature and in assessing the current methodological limitations. However, the implementation of this study as a registered-report is still suboptimal, my main concerns are outlined below.

We are happy that you like the topic and the concept of our study. Thank you for pointing out methodological limitations of our planned study. We tried to thoroughly check and solve these issues and outline these changes in detail below.

1. The existing section "How could absolute magnitude affect…" left me wondering whether this is all really necessary, or whether this part could be shortened focusing on Table 1 which seems the one that readers can easily link to the rest of the manuscript.

   We have shortened the section and made it more concise. Nevertheless, wish to keep the elaboration of the six scenarios in form of a text as well, because these scenarios are essential for the rationale of our study and the text might be easier and more enduring to understand for some readers than only the graphical illustration of the scenarios and Table 1.

2. Statistical power. The authors opted for d=0.2 as minimal effect size of interest and explained how estimating this effect based on the existing literature could be biased by various factors. As a reference, it would be anyway useful to report what the typical effect size in this literature is. What was the original effect size in the studies they aim to replicate? Also, when reporting the a priori calculation, the authors refer to a specific standard deviation but do not report any value - please add.

   Thanks for the helpful suggestions. We inserted the exact standard deviations reported by Fias et al.'s (1996). Moreover, we calculated the effect size in terms of Cohen's *d* for the study by Fias et al., which is approximately $d = 0.16$, and report it in the section *Hints towards AMdependency of the SNARC effect*. Note that the standard deviation in their study was rather large and that we observed three times smaller standard deviations in our previous color judgment experiments. A smaller standard deviation, which we expect, leads to a larger effect size.

   As outlined in the Registered Report, it is not possible to calculate the effect size for the study by Dehaene et al. (1993) due to the lack of reported standard deviations. Nevertheless, they observed a descriptively much larger slope difference, so the effect

size was likely greater in their study than in Fias et al.'s study, if they had not much larger standard deviations in their data.

Importantly, we have added to the paragraph *Statistical power considerations and sample size determination* of our Registered Report that importantly, "in the two original studies, the symmetric confidence intervals for these estimates must also include at least the double slope difference and effect size due to their non-significance. Hence, in case of AMdependency of the strength of the SNARC effect, the true effect size might in fact be larger than d = 0.16." This way, readers get a feeling for plausible effect sizes, but we remind them that these effects remained non-significant in previous studies and that these estimates were therefore very imprecise.

3. Participants. The authors report only a minimal age (18) as a requirement. Since the experiments measure reaction times, for which age differences might exist, wouldn't it be more sensible to also add a maximal age? What's the aim of giving not just full but also partial compensation?

   Thank you for this suggestion. There might be age differences for RTs, which typically correlate with the size of the SNARC effect. We have inserted a maximal age of 40 years in the newer version of our Registered Report. If Reviewer 2 or the Editor has other ideas on the age limit, we are happy to incorporate them.

   Apart from full participation, we need to compensate for partial participation, because this is required by our ethics committee. They advise us that participants' time and effort should still be valued if they only complete a part of the study (e.g., if their internet connection breaks down), although we cannot use partial datasets. Therefore, we will proportionally pay them, e.g., if participants only complete the first half of the study, they will get half of the payment.

4. Procedure. The experiment will be implemented in the Wextor online platform. Since the expected effects are very small, do the authors have any information regarding the measurement accuracy of this tool (e.g., compared to lab-based experiments?)

   The experiments have been set up with WEXTOR, which is a tool that helps researchers to create experiments in HTML and JavaScript in a guided way (demo versions are available at https://luk.uni-konstanz.de/numcog_3/?demo&e1 and https://luk.uni-konstanz.de/numcog_3/?demo&e2). Therefore, the measurement accuracy is not dependent on WEXTOR, but instead on the programming and on the web browser participants use to carry out the experiment. As outlined in Garaizar & Reips (2019), while browsers are designed to provide the highest speed to increase the responsiveness of web applications, behavioral researchers need high precision and accuracy when presenting stimuli and recording responses. Therefore, all materials for our experiments (such as instructions, number stimuli, questions, etc.) are downloaded at the very beginning when participants access the website. This way, the browser can subsequently render the experiment while avoiding large delays in presenting stimuli and recording responses.

5. Quality check. The authors report a seriousness check that will be used prior to the beginning of the procedure, and a self-assessment to be filled right afterwards (e.g., the participants will rate the condition in which the experiment took place etc). However, they do not report any concrete quality check to assess correct implementation of the procedure and participant's compliance with instructions. In line with this, the authors do not appear

to have implemented any positive control. These aspects need to be carefully addressed for any registered report, especially in online procedures.

Thanks for pointing this out, we have inserted a new subsection with the header *Positive controls and manipulation checks* to our manuscript:

"To control the data quality in our study, we have implemented a seriousness check (Aust et al., Reips, 2009, review in Reips, 2021) as well as a self-assessment of noise, distractions, and other difficulties. To make sure that we will only analyze trials that reflect mental processes in correctly executed parity judgment, we will exclude incorrectly answered trials and trim RTs (as described in the data preprocessing pipeline). Also, we will exclude full datasets of participants with less than 75% valid trials to only build our results on participants who have understood and followed the task instructions. Moreover, we ask participants whether they complied with the instructions to use their left and right index fingers for the left and right response keys, respectively.

Last, we will check for the presence of the Odd Effect (Hines, 1990; i.e., overall faster reactions to even than to odd numbers, irrespective of the response side). The Odd Effect is quite robust in the parity judgment task, but independent from the SNARC effect (as it is independent from number magnitude and from its mapping onto space and only considers parity). Therefore, we can consider its investigation as a manipulation check, and in case of its presence we will have a positive control for our experiment. For this, we will subtract the average RT for even numbers from the average RT for odd numbers per participant and test the differences (one per participant) against zero in a Bayesian one-sample *t*-test (with positive estimates indicating the Odd Effect)."

6. Demographic questions. What's the rationale behind allowing the "I prefer not to answer" option? It seems rather essential to collect complete answers from all participants. Also, why explicitly use the term "finger counting habit" in these questions? This might seem rather obscure to the participants.

   It seems of course essential to collect complete answers, but at the same time, an even higher priority for us is to collect high-quality answers (i.e., honest answers). In other words, as we have added to the revised manuscript, we implemented this option "to respect some participants' unwillingness to share information with us and to not force them to choose any option that might not reflect the truth (Jenadeleh et al., 2023; Stieger et al., 2007). Note that in earlier studies, only very few participants chose this option in any of the above-mentioned questions.", so we are not afraid to miss much information here anyway. Also, providing the option to choose "I prefer not to answer" is necessary to comply with the demands of our ethics committee.

   Thanks for pointing out the obscure terminology in our demographic questions. We will change the item "Your finger counting habits" to "How do you count with your fingers?" and the item "Stability of your finger counting habits" to "How often do you start counting with your fingers with this hand?" in the demo versions and in the real experiment as soon as possible.

7. Response keys. The phrasing here is rather obscure to me. Why allow the participants to use any other key than the two that were assigned by default? Especially if no check is put in place, e.g. how will they check that the distance between the two keys is optimal?

We implemented the option to use other response keys than the default (D and K as left and right response keys) in case participants use an unconventional type of keyboard, where D and K are not in the same row or not approximately one hand width apart from each other. With this technical improvement of our experiment, we can avoid potential confounds (e.g., if the left response key is above the right response key, there would be a confound between left-right directionality and up-down directionality).

We cannot guarantee that participants use the index finger of their left hand for the left response key and the index finger of their right hand for the right response key. Controlling for this is very difficult in an online study, no matter whether participants have the option to change the default response keys or not (as a side note, it is also not entirely feasible to ensure it in lab studies, as this would require experimenter constantly looking at participants' hands, and this would cause unnecessary stress for the participant, not even mentioning group administration of the parity judgment task). However, we instruct them to use both hands, and we do not see any reason for then not complying with these instructions. Given the distance of the two response keys on the keyboard (namely, approximately one hand with), it should even be uncomfortable for participants to use just one hand for both response keys. Nevertheless, we will add a control question at the end of the experiments as soon as possible, where we ask participants whether they have used their left and right index fingers.

8. Dropout rates. The authors plan to further investigate the reasons for dropouts, but it's unclear how the results of this analysis will affect subsequent analyses (e.g., in case they show significantly different dropout rates in some conditions?)

   Given that we do not have different conditions in a terms of experimental manipulations, but only different block orders resulting from counterbalancing in our within-subjects design, we do not expect different dropout rates. However, because MARC incongruent blocks (i.e., even and odd numbers being assigned to the left and right response keys, respectively; cf. Nuerk et al., 2004, for the MARC effect) might be cognitively slightly more demanding for some participants than the MARC congruent blocks (i.e., odd and even numbers being assigned to the left and right response keys, respectively), the possibility remains that dropout rates are higher when participants start with the former than with the latter blocks. Hence, it is hard to make any clear predictions for differential dropout rates between counterbalanced block orders and results will not affect subsequent analyses. Therefore, we will run the dropout analysis as an exploratory analysis and have removed it from the Registered Report.

9. Statistical approach. The authors aim to combine null-hypothesis testing with estimation of bayes factors. I assume the former was chosen because of earlier studies, however since the main analyses will employ t-tests why not opt directly for a full bayesian approach? Combining the two approaches always appears rather complex to manage in a registered report - especially when outline the interpretations based on different outcomes. A plus of opting for a bayesian approach is that it would allow the authors to use sequential analyses for a more efficient recruitment and sampling plan (e.g., Schönbrodt, F.D., Wagenmakers, EJ. Bayes factor design analysis: Planning for compelling evidence. Psychon Bull Rev 25, 128–142 (2018))

   Thanks for this helpful comment. We have decided to employ only the Bayesian approach instead of combining it with the frequentist approach. We appreciate your

suggestion to use sequential analyses with optional stopping and decided to go for the "Sequential Bayes factor with maximal n" (SBF+maxN) approach as suggested by Schönbrodt & Wagenmakers (2018).

We have revised the section Statistical power considerations and sample size determination in our manuscript and explained the SBF+maxN approach there. Also, we have revised the Study Design Table. Namely, we have set the minimum sample size to 200 participants and will analyze the data after each 20 new datasets until we reach moderate evidence against or for AMdependency of the strength of the SNARC effect (i.e., threshold of $BF_{10} > 3$ or $BF_{10} < 1/3$ for or against Hypothesis 3). We have chosen to make the optional stopping dependent on Hypothesis 3 because it is the most crucial hypothesis in our study. Moreover, a sample that is large enough to find moderate evidence against or for Hypothesis 3 is most probably also large enough to find at least moderate evidence for the SNARC effect in different ranges (Hypothesis 1) and for both RMdependency and AMdependency of the number mapping on the MNL (Hypothesis 2). We have defined a maximum sample size of 700 participants, which is the sample size that we have determined to be necessary for detecting an effect of Cohen's $d = 0.15$) with power of .90 when using the rather large standard deviations observed by Fias et al. (1996). Note that although the term "moderate evidence" does not sound very convincing, the respective BF_{10} threshold of 3 corresponds to p-values below .01, so that we consider this to be a rather conservative and adequate threshold.

10. Analyses. The authors report three types of analyses: main, follow-up and exploratory. However, by definition a Stage 1 submission that will later become a pre-registration, cannot include exploratory analyses. While they could be generically referred to in the analysis plan, they cannot be outlined in detail and included in the design planner - otherwise they'd be pre-registered as well. I'm more uncertain regarding the follow-up analyses, which have a more nuanced status - I invite the authors to reconsider whether these analyses should be pre-registered or not.

We have removed the entire paragraphs 'Follow-up analysis' and 'Exploratory data analysis' from the manuscript.