Dear Prof. McIntosh,

Thank you for giving us the opportunity to submit another revised version of our Registered Report. Just as in the first review round, we found your comments and suggestions very helpful. Below, you can find our replies to your comments. We revised the manuscript accordingly and highlighted all changes that we have implemented in the second review round.

Kind regards,
Lilly Roth (on behalf of all authors)

Thank you for your careful work addressing review comments for this Stage 1 RR. The revisions have been evaluated by one of the original reviewers, who is happy with the changes. (The other original reviewer was not available at this time.)

We are happy to hear that the reviewer is satisfied with the changes.

I have looked over the paper myself, and think it is considerably improved, but there are a few more issues that I would like you to give attention to before IPA is issued for this experiment. You are not obliged to follow any suggestions made, but should provide a rationale for the course of action that you decide.

1) "Power analysis". In developing you study plan, you present both frequentist power analyses (for prior studies) and power analyses developed within the Bayesian framework for your planned study. Given that you seem to be applying a criterion threshold BF (>3) to support a binary claim, it may be legitimate to talk about 'power', but I think it is nonetheless potentially confusing for you to use the same language of 'power' to apply both to frequentist and Bayesian approaches.
One key reason is that power concerns only the probability to detect true effects (of a given size) when present. But your Bayesian analysis is not only asking this question, but is configured also to return evidence in favour of the null hypothesis (BF < 1/3). Thus, you state in the abstract that "... a power of .90 for detecting moderate evidence (Bayes Factor above 3 or below 1/3)", but actually your sample size planning seems to be predicated only on sensitivity to a true effect when present, without considering your sensitivity to the null when the null is true. A small tweak of your code would allow you to make more complete statements about the probability of your Bayesian analysis to return sensitive evidence for H1 or H0, and the rates of misleading evidence (using the language of Bayes Factor Design Analysis).

Thanks for pointing out this important confusion between Bayesian and frequentist analyses. We really appreciate your explanations and are glad that you raised these concerns.

In a revised version of our RMarkdown script, we calculate the probability of obtaining evidence for both H1 (given a specific effect size) and H0 (i.e., when the effect size = 0). We found that we only need 180 participants to find at least moderate evidence against the effect in case it is truly absent. This is much less than 800 participants that are required to find at least moderate evidence for a true underlying effect of the minimally relevant effect size of Cohen’s *d* = 0.15. For the SBF+maxN approach, we will go for the larger of these two required sample sizes as a maximal sample size, which is 800.

We have adapted the wording in the manuscript, in the *Study Design Table*, and in the RMarkdown script accordingly: We now talk about the "probability of finding at least moderate evidence" instead of the "statistical power for rejecting the null hypothesis" in case a true underlying effect of at least the minimally relevant size exists. This will avoid confusions between the Bayesian and the frequentist framework. When explaining the power-determination analysis and effect-size sensitivity approach for the two seminal studies (Dehaene et al., 1993; Fias et al., 1996), we kept the frequentist terminology.

Importantly, we have renamed the section *Statistical power considerations and sample size determination* of our *Method* part. It is now called *Sample size considerations*. We have added an explanation of how we determined the maximal sample size (pages 18 and 19): "The sample size of 800 participants is required for a proportion of at least .90 Bayesian *t*-tests to yield a BF_{10} above 3, when 5000 samples of SNARC slope differences are randomly drawn from a normal distribution around the minimally relevant effect size of $d = 0.15$ are simulated (for a similar approach, see Kelter, 2021). Following the same procedure, we found that the sample would need to consist of 180 participants to ensure a probability of .90 for finding at least moderate evidence against a truly absent effect (i.e., BF_{10} below 1/3 for $d = 0$, according to Dienes, 2021). Note that the sample size of 180 is smaller than the initial sample size of 200 that will be collected in the SBF+maxN approach.".

As a final note, we have added the following sentence to give readers who are unfamiliar with the Bayesian equivalent of frequentist power simulations an idea of how required sample sizes relate to each other (page 19): "While in the frequentist framework, low error type II rates (and high statistical power) need to be achieved, in the Bayesian framework, low rates of misleading evidence (and a high probability of finding evidence for a true underlying effect) need to be ensured. To achieve the same probability for error type II and misleading evidence, Bayesian *t*-tests (using the default *r*-scale of 0.707 as uninformed prior in the Cauchy distribution) require larger samples as compared to frequentist *t*-tests (Kelter, 2021)."

2) At present, your sample size is predicated on the smallest effect size of interest for H3, and the tests of H1 and H2 inherit their sensitivity from this design. Strictly speaking, this means that your experimental design has not been shown to be adequate to test H1 and H2, whereas the RR format required that you demonstrate your required standard of evidence for all hypotheses. Given that your SESOI for H3 is so small, it would seem to be a small extra step for you to make the (easy) argument that an effect size smaller than this would be similarly uninteresting for H1 and H2, which would then allow you to assert the required level of sensitivity for all hypotheses.

Thank you for pointing this out and suggesting a straightforward and adequate solution. As you say, we use the same statistical tests (namely, Bayesian one-sample *t*-tests or paired *t*-tests) in all hypotheses and an effect smaller than $d = 0.15$ would not be meaningful for Hypotheses 1 and 2 either. Therefore, we added to the paragraph *Statistical power considerations and sample size determination* in the *Method* part on page 18: "This sample size estimation is also valid for testing Hypotheses 1 and 2, which require one-sample *t*-tests. The reason is that an effect smaller than $d = 0.15$ would not be meaningful for the SNARC effect in the lower (Hypothesis 1a) or higher

(Hypothesis 1b) number range or for RMdependency (Hypothesis 2a) and AMdependency (Hypothesis 2b) of the number mapping on the MNL either. Similarly, the chosen maximal sample size should be large enough to find at least moderate evidence in case these hypotheses are false."

Similarly, we adapted this argument for the columns *Sampling plan* and *Rationale for deciding the sensitivity of the test* in the *Study Design Table*.

3) However, for your experiment to really have the required level of sensitivity for all hypotheses, then your stopping rule cannot be based on a sensitive outcome for one hypothesis only - you could only stop the experiment (prior to n-max) if a sensitive BF were found for all three hypotheses. If your plan is to terminate the experiment based on H3 only, then your plan does not have the desired level of sensitivity for H1 and H2, and you would need to relegate these hypotheses to secondary, exploratory status (i.e. remove them from the Stage 1 plan).

Thank you for this comment. We have adapted the text in the paragraph *Statistical power considerations and sample size determination* in the *Method* part (page 19), so that the stopping rule is based on a sensitive outcome for all three hypotheses now: "We use moderate evidence in favor of all hypotheses ($BF_{10} > 3$) or against them ($BF_{10} < 1/3$) as thresholds. More precisely, for each experiment, we will first recruit 200 participants and compute the BF_{10} for the SNARC effect in lower (Hypothesis 1a) and higher (Hypothesis 1b) number ranges, for the shift of critical small/large numbers in both relative (Hypothesis 2a) and absolute (Hypothesis 2b) terms towards the left/right, respectively, and for the SNARC slope difference between ranges (Hypothesis 3). As long as the BF_{10} does not reach any of the two thresholds for all hypotheses, we want to collect another 20 datasets and recalculate the BF_{10}. If no threshold is reached with our maximal sample size of 800 participants (that is required for obtaining at least moderate evidence for a true underlying minimally relevant effect with a probability of at least .90, as explained above), we will stop the sequential recruiting of participants in any case."

We have similarly adapted the description of the stopping rule in the *Study Design Table*.

4) You have now added the Odd Effect as a positive control/manipulation check. However, your logical chain here simply states that it is a robust effect in parity judgements, and that you expect to find it and will be surprised if you find evidence against it (you do not state what will happen if the BF is insensitive). This does not constitute a meaningful manipulation check, because it does not seem to have any implications for your main hypothesis tests. Normally, a manipulation check is an effect that should definitely be present in the data so that, if it is not found, there is evidence that your task has not worked as intended. Normally, when a manipulation check is failed, then the conclusion is that the experiment is deemed incapable of returning a clear answer on the experimental hypotheses. For this reason, the manipulation check is normally first in the list of inferential tests, to establish the adequacy of the task to the question. Like other inferential tests, it requires a power/sensitivity analysis. If the Odd Effect has this status, then you need to make this clear. If it does not, then it is not a manipulation check.

Thank you for this helpful comment. In the end, we deleted the Odd effect as a manipulation check from the manuscript. It is a robust effect, but its presence is not a prerequisite for investigating our main hypotheses (i.e., RMdependency and AMdependency of the number mapping on the MNL, and AMdependency of the SNARC effect). Instead, as you suggested in your next point, we decided to use the presence of a SNARC effect in the lower number range (Hypothesis 1a) as a manipulation check, see details below (in response to your fifth comment).

5) On the other hand, your H1 is a check that the SNARC effect is present in all number ranges. This seems much more to me like a conventional (and relevant) manipulation check, and yet you simply state that you will be surprised if you don't find it in all ranges, but do not indicate that this would limit your ability to test further experimental hypotheses in any way. I would at least have thought that finding the SNARC effect in a given range was a requirement for testing the other hypotheses *with respect to that range* (i.e., any tests in which that range is involved for H2 or H3). If this is not the case, then it would seem that your experiment has been configured such that you could be making claims about the range dependency of the SNARC effect even if your data showed no evidence of SNARC effects *per se*. I realise that this outcome is unlikely, but it is the logical coherence of your analysis plan that is at stake.

Thank you for your valuable suggestion and for all explanations. In our revised manuscript, we now use the SNARC effect in the lower number range as a manipulation check for each experiment.

Moreover, we have changed parts of our *Data analysis* part accordingly (pages 26 and 27): "First, we will test the presence of the SNARC effect on group level in both number ranges separately in each experiment (Hypothesis 1). As described in the introduction, the SNARC effect seems to be stronger in the lower than in the higher number range in terms of a more negative slope. As the SNARC effect is very robust especially for lower ranges and possibly stronger than in higher ranges (see Hypothesis 3), the SNARC effect in lower ranges (Hypothesis 1a) will be used as manipulation check and prerequisite for following investigations (Hypotheses 1b, 2 and 3). […] Evidence for the SNARC effect in all ranges would replicate findings from the two seminal studies by Dehaene et al. (1993) and Fias et al. (1996). The lack of conclusive evidence as regards the SNARC effect in the lower ranges (Hypothesis 1a) with our maximal sample of 800 participants or even evidence against it is highly unlikely. Evidence against the SNARC effect in the higher ranges (Hypothesis 1b) combined with evidence for the SNARC effect in the lower ranges (Hypothesis 1a) would provide support for AMdependency of the strength of the SNARC effect (Hypothesis 3)."

We also framed the lower-number-range part of *Hypothesis 1* (which is now called *Hypothesis 1a*) as a manipulation check in our *Study Design Table* and stated that not finding the SNARC effect in the lower number range would have the consequence that we will not test any further hypotheses for the respective experiment.

6) The manuscript is long and complex. There are no word limits at PCI-RR, but with the aim of future publication, I would strongly encourage you to try to be more concise wherever possible (obviously, without omitting any essential material).

Thanks for your advice. We have shortened the manuscript and made it less complex by taking out the detailed elaboration of the six scenarios regarding RMdependency and AMdependency of both the number mapping on the MNL and the strength of the SNARC effect (pages 10 to 15). We have posted the part of the manuscript that contained the six scenarios together with their illustrations in figures and tables on OSF as supplementary material and linked to it in the manuscript, so that interested readers still have access to this part.

As a replacement, we have shortly summarized the core information which is a prerequisite to understand our three hypotheses in the paragraph *The current study* (page 15): “Crucially, in contrast to previous literature about the flexibility of the SNARC effect, we will differentiate between two concepts that can be affected by RMdependency and AMdependency:

(i) On the one hand, the number mapping on the MNL (e.g., dRT for number 4) may be different depending on the experimental setup. In our setup, it can be RMdependent (i.e., depending on the position on the used range, e.g., position 5 for range 0 – 5, or 1 for range 4 – 9), AMdependent (i.e., depending on the magnitude, e.g., 4), or both at the same time.

(ii) On the other hand, the strength of the SNARC effect relies on the relative increase of right-hand advantage per increase in magnitude (i.e., the steepness of the SNARC slopes, e.g., -5 ms per number or -10 ms per number) and these slopes can differ between ranges.

For a more detailed but rather complex elaboration of six possible scenarios combining different parameters of (i) and (ii), see Figures S1 and S2 in our Supplementary Material (https://doi.org/10.17605/OSF.IO/Z43PM).”

Hypotheses 1, 2, and 3 are still part of our manuscript and still meaningful, even without background information on the six possible scenarios. We have split Hypotheses 1 and 2 into parts (a) and (b) and refer to them more precisely in the data analysis section, so that readers can more easily follow. To make hypotheses well understandable for readers who do not look at our Supplementary Material, we have added some more explanations to them (page 17):

“1. A SNARC effect in both (a) the lower and (b) the higher number ranges in each experiment. (a) The SNARC effect in the lower range serves as a manipulation check and is considered as a prerequisite for testing Hypotheses 2 and 3 in the respective experiment. Both (a) and (b) aim at replicating the results by Dehaene et al. (1993) and Fias et al. (1996).

2. Both (a) RMdependency and (b) AMdependency of the number mapping on the MNL, such that small/large numbers in relative and absolute terms are shifted towards the left/right, respectively. (a) RMdependency would be reflected by dRTs for the same critical numbers (i.e., 4 and 5) differing between ranges, showing that the MNL adapts flexibly and relative to the range. (b) AMdependency would be reflected by dRTs for these critical numbers being equal between ranges, and by dRTs for the smallest number in each range (Experiment 1: 0 in the 0 – 5 range vs. 4 in the 4 – 9 range; Experiment 2: 1 in the 1 – 5 range [excluding 3] vs. 4 in the 4 – 8 range [excluding 6]) differing between ranges, AMdependency would mean that small/large numbers are shifted to the left/right on the MNL, although they are exactly on the same position within their range, but differ in terms of absolute magnitude.

3. AMdependency of the strength of the SNARC effect, such that it is stronger in the lower than in the higher ranges. This would be reflected by steeper (i.e., more negative) SNARC slopes in the lower than in the higher ranges, which was

descriptively observed in the two seminal studies by Dehaene et al. (1993) and Fias et al. (1996)."

Our revised manuscript is now shorter and more straightforward, especially for readers who are not familiar with the SNARC effect and with assumptions about its flexibility that we aim to thoroughly test in the study.

7) The new title is rather unwieldy: "“One and only SNARC? How Flexible are The Flexibility of Spatial-Numerical Associations? A Registered Report on the SNARC's Range Dependency".
Should the first part be: "One and only one SNARC?"? The second and third parts seem to be two alternate sub-titles; there should probably be one and only one sub-title.

We have changed the title to "One and only SNARC? A Registered Report on the SNARC Effect's Range Dependency", so that it is less unwieldy and more concise, and we indeed have one and only one subtitle.

I hope that these comments are helpful. You have done a lot of good work to sharpen this plan up, but I think that a little further sharpening and streamlining is required before IPA.

Best wishes,
Rob McIntosh

Thank you so much, your comments were very helpful!

We have added a few more sentences to the *Procedure* paragraph of the revised version of our manuscript because we have noticed that it did not contain the exact instructions for participants (although they can be seen when readers check out the demo versions of our experiments that we have made publicly available). In the revised version of the manuscript, we have inserted the exact wording of our instructions for an exemplary experimental block (page 24):
"In our experiment, your task is to distinguish the parity of numbers, that is, to decide whether a number is even or odd. For this, please place the index finger of your left hand on the [D] key and the index finger of your right hand on the [K] key on your keyboard. In each run, a black square will appear in the center of the screen. Please look at this square. It will soon be replaced by either an even or an odd number. If the number is even (0, 2, 4), press [D]. If the number is odd (1, 3, 5), press [K]. Please answer as quickly and as accurately as possible."

We are looking forward to hearing back from you again.

Best wishes,
Lilly Roth (on behalf of all authors)