**PCI Registered Reports #794**

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

**Stage-1 Submission, Version 2**

Dear Mario,

Thank you for giving us the opportunity to submit a revised version of our registered report manuscript with the title "Shape of SNARC: How task-dependent are Spatial-Numerical Associations? A highly powered online experiment". We are grateful for the three reviewers' valuable feedback and appreciate the helpful comments.

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

Best wishes,
Lilly (on behalf of all authors)

Firstly, I apologise for the extended delay due to the challenge of finding three qualified and expert reviewers during this period. I want to express my sincere gratitude to the three reviewers (reviewer 3 is Peter Wühr) for their valuable comments.

All three reviewers liked the proposal and agreed that this study will provide important insights into the literature, and I agree with their assessment.

We are glad that three reviewers and you like our submitted stage-1 registered report. We thank all four of you for your very constructive feedback.

As you will see, the reviewers provided several practical suggestions. In particular, the first reviewer requested clearer hypotheses and more discussion on the MC-SNARC and PJ-SNARC correlation, raising concerns about the reliability due to the low consistency of SNARC effects. The second reviewer suggested focusing the introduction more on the study's main goal and also commented on the proposed Bayesian analyses. The third reviewer asked for a clearer explanation of the dual-route model and proposed additional, exploratory analyses to explore whether SNARC effects change over time, particularly between tasks.

We agree that our theoretical considerations were too extensive and did not necessarily help deriving differing hypotheses in the original manuscript. Therefore, we followed the recommendation to focus more on the methodological importance of the paper and refrained from some of the theoretical considerations. Specifically, we shortened the part about the correlation between the MC-SNARC and the PJ-SNARC by not deriving predictions from SNARC theories any longer but instead mentioning the reliability issues as pointed out by Reviewer 1. We also cut on some SNARC theories, for instance the cognitive-control account for the SNARC effect by Zhang et al. (2022) and the dual-

route model by Gevers, Ratinckx, et al. (2006), and only mention them briefly in parentheses. This also resolves the concerns about our predictions derived from the dual-route model that were raised by Reviewer 3. We thereby make the entire Introduction more concise, as requested by Reviewer 2. Moreover, we added and modified some analyses as suggested by Reviewers 2 and 3.

On a side note, we renamed "manipulation checks" and call them "replication checks" in the revised manuscript. The first reason for renaming them was that the abbreviation MC referring to "magnitude classification" might have been misinterpreted by readers as standing for "manipulation check". The second reason for renaming them was that only the first "manipulation check" (i.e., a SNARC effect in both MC and PJ with the standard analysis of a continuous linear regression) will be used as a positive control. We did not plan to use any "replication check" other than the first as a prerequisite for testing all main confirmatory analyses, so the terminology "manipulation checks" was not suitable.

Finally, to give readers a better overview of all planned analyses, after the list of replication checks and hypotheses we wish to test, we also provide a summary of all questions we aim to answer in an exploratory manner (pages 23 and 24):

"Moreover, we will explore whether the following observations can be made (without directional predictions):

1. task-order effects on both (a) the MC-SNARC and (b) the PJ-SNARC;
2. a good model fit when including both continuous and categorical magnitude predictors for (a) the MC-SNARC or for (b) the PJ-SNARC, indicating a mixed shape of the SNARC effect (see Panel C in Figure 2);
3. compatibility-order effects on (a) the MC-SNARC (SNARC slopes in Conditions 1 and 3 versus Conditions 2 and 4) or on (b) the PJ-MARC (MARC slopes in Conditions 1 and 3 versus Conditions 2 and 4);
4. a shape difference of (a) the MC-SNARC or for (b) the PJ-SNARC between earlier and later phases within each task;
5. a correlation between the categorical MC-SNARC slopes and the continuous PJ-SNARC slopes."

I would encourage you to resubmit the proposal by taking into account the suggestions made by the three reviewers.

by Mario Dalmaso, 12 Sep 2024 14:31

Review by anonymous reviewer 1, 17 Jul 2024 09:57

This pre-registered study aims to explore significant research questions with appropriate and fully reproducible methodologies. While the SNARC effect is a well-known phenomenon studied using various approaches and perspectives, some critical aspects remain underexplored. These include the exact 'shape' of the effect, specifically which mathematical model best captures the relationship between numerical magnitude and dRTs, and the relationship between the SNARC effect obtained through parity judgment and magnitude comparison tasks. This study aims to address these two important issues, potentially providing a valuable contribution to the literature.

The procedure for sample size determination and the analysis plan are robust, requiring no further suggestions for improvement. I would like to commend the authors for the robust and detailed description of the criteria for sample size determination and the analysis plan. This thoroughness sets a valuable example for enhancing the methodological rigor of studies in this research field. However, there are some suggestions to improve the clarity of the hypotheses definition.

> We are happy that the reviewer appreciates our efforts to make this study as transparent as possible. In our view, sharing materials to make research studies understandable and reproducible is crucial.

1. On page 5, a reference to the concept of 'primitive' is made without a specific definition. In this context, 'primitive' seems to imply an 'automatically activated process.' However, a precise definition of this concept is needed to enhance clarity.

> We thank the reviewer for spotting this. We have deleted the term "primitive" from some early spots in the paper, and we have now properly introduced single-digit Arabic numbers as primitives (page 5) by explaining that "their meaning can be holistically retrieved from memory without further processing" in Western cultures (cf. Tzelgov et al., 2015).

2. On page 6, at the end of the section entitled 'Relevance of number magnitude and parity,' a discussion of the distance effect is notably absent.

> The reviewer is right, so we introduce the Numerical Distance Effect earlier (pages 6 to 7) already in the revised manuscript, right after introducing the Numerical Size Effect: "Moreover, RTs increase with increasing numerical distance between the stimulus and the reference number in MC, which is referred to as the Numerical Distance Effect (Gevers, Verguts, et al., 2006). We expect the might effect to arise in MC (Replication Check 4), but it cannot arise in PJ because there is no criterion that numbers are compared to)."

3. I have several concerns regarding the content of the 'Correlation between MC- and PJ-SNARC' section:

- Before discussing any predictions about the possible correlation between the MC- and PJ-SNARC, it should be explicitly acknowledged that previous studies have shown that the test-retest reliability of the SNARC effect is quite poor for both the MC task (Hedge et al., 2018) and the PJ task (Viarouge et al., 2014). As explained by Hedge et al. (2018), this low reliability likely has nothing to do with the stability or instability of the cognitive processes underlying the overt effect (i.e., the SNARC), but rather is related to the low inter-individual variability of the effect itself. In simple terms, low inter-individual variability imposes an upper limit on the magnitude of test-retest reliability.

- The correlation between PJ- and MC-SNARC can be expected to be, at most, as large as the test-retest reliability observed for a single task. However, it seems safer to predict that it might be even smaller due to superficial differences between the tasks. In light of this, it does not seem legitimate to test different theoretical predictions based on the observed correlation coefficient. For instance, a high correlation appears implausible regardless of the correctness of the MNL theory, and a low correlation is expected regardless of whether the dual-route or the WM accounts are correct or not.

> We appreciate the constructive feedback and agree with the limitations you name. Hence, we have modified the beginning of the section (pages 12 to 14):
> "After having described the similarities and differences of MC and PJ, the question arises whether the MC-SNARC and the PJ-SNARC are correlated. However, both factors at the construct level of SNAs and at the operational level of the two tasks might lead to a null correlation. First, there seem to be high fluctuations in the SNARC effect over time (Roth, Jordan, et al., 2024) that limit the maximum correlation that can be detected. Second, the test-retest reliability of the SNARC effect has been found to be poor for both MC and PJ (correlations $.22 < r < .41$; Cipora & Göbel, 2013; Georges et al., 2013; Hedge et al., 2018; Viarouge et al., 2014). The lower the test-retest reliabilities of the MC-SNARC and PJ-SNARC, the lower is also the maximally observable correlation between the two effects. Third, the split-half reliability of the SNARC effect has been found to be poor for PJ at least in some studies (correlations $.43 < r < .96$; for an overview, see Cipora, van Dijck, et al., 2019, Table 1 there). To conclude, both properties of the SNA construct and its operationalization in experimental tasks influence whether a correlation between the MC-SNARC and the PJ-SNARC will be found. Possible reasons for a null finding could be low intraindividual stability, low reliability, or low internal consistency, whereas a high correlation between the MC-SNARC and the PJ-SNARC would lead to the conclusion that both MC and PJ reliably measure the same underlying theoretical construct."

- The authors briefly acknowledge this potential issue at the end of the section. This is confusing because if the authors are aware that the correlation between MC- and PJ-SNARC is likely to be small, then predicting that it might be large or moderate is illogical. In other words, comparing different theories based on the observed correlation coefficient seems impossible from any perspective.

- Apart from these crucial issues, I have concerns about the consistency of some theory-driven predictions regarding the correlation coefficient. For instance, the authors suggest that, according to the hypothesis that the MC- and PJ-SNARC are related to different WM systems

(verbal or visuospatial), no correlation is expected. However, it seems reasonable to expect some correlation between verbal and visuospatial WM capacities. At the very least, some positive correlation is expected due to the superficial similarities between the tasks, as some general cognitive skills are likely involved in both tasks.

- In light of these concerns, I think that this section requires substantial revision. I suggest that the authors focus on the methodological issues related to interpreting the correlation between MC- and PJ-SNARC, refraining from making any theory-based predictions. Alternatively, if predictions about the correlation coefficient are discussed, it should be made clear that methodological issues render the comparison practically impossible.

> We sincerely thank the reviewer for their well-justified argumentation. We followed their advice to refrain from theory-based predictions about a potential correlation between the MC-SNARC and the PJ-SNARC, and we focus on methodological issues regarding the exploration of a potential correlation.
> In the revised manuscript, we do not make any predictions about the potential correlation. That is, we deleted all three theories (the MNL account, the dual-route model, and the WM account) from this part of the manuscript. Instead, we now write (page 13): "Possible reasons for a null finding could be low intraindividual stability, low reliability, or low internal consistency, whereas a high correlation between the MC-SNARC and the PJ-SNARC would lead to the conclusion that both MC and PJ reliably measure the same underlying theoretical construct."
> We conclude that "to be able to detect a potential correlation between the MC-SNARC and the PJ-SNARC, we will administer both tasks to a large sample in a within-subjects design." (page 15)

- Additionally, I recommend the authors discuss an important difference between the present study and previous studies on the test-retest reliability of the SNARC effect. Unlike previous studies, this study does not include a 'washing-out' period between the two tasks. This procedural difference should be adequately considered when interpreting the results.

> We thank the reviewer for this remark. It is true that our planned study does not include a 'washing-out' period between MC and PJ. However, we wish to note that assessing test-retest reliability is not the aim of our study and instead of testing the same task twice, we wish to test two different SNARC tasks. We intentionally let participants complete the second task right after the first to be able to investigate spillover/transfer effects. Precisely, this procedure in combination with the conterbalancing of task order between subjects permits us to compare the strength of the MC-SNARC and of the PJ-SNARC between task orders.
>
> We have added the following sentences to the Introduction, where we explain block order effects (pages 8 and 9):
> "Bae et al. (2013) and Bulut, Çetinkaya, et al. (2024) demonstrated that the response-to-key assignment in MC influences the SNARC effect in subsequently measured PJ. Specifically, they found a regular left-to-right number mapping in PJ after a SNARC-compatible MC block (i.e., small-left and large-right), but a reversed right-to-left number mapping in PJ after a SNARC-incompatible MC block (large-left and small-

right). However, in these studies, participants were assigned to only one of two possible response-to-key assignments for MC. Hence, habituation or practice that spilled over from MC to PJ was unidirectional, and furthermore, no MC-SNARC could be determined."

- Lastly, it appears that the discussion in this section does not translate into a specific analysis plan (see pages 19-20). Therefore, clarification is needed regarding the exact role of the correlation between PJ- and MC-SNARC within the broader analysis plan.

In the revised manuscript, we mention that our within-subjects design "enables us to test the correlation between the MC-SNARC and the PJ-SNARC in an exploratory analysis with high statistical power." (page 15)
We do not derive any theory-driven prediction about whether or not we will observe a correlation, so we do not include any prediction in our replication checks or hypotheses. Instead, we state here that we will run an exploratory analysis, and readers can more easily find the respective analysis in the Method section (page 38): "We will calculate Pearson's correlation between the categorical MC-SNARC slopes and the continuous PJ-SNARC slopes (Exploratory 5). We will run a two-sided Bayesian Pearson correlation test to see whether the spatial mapping of number magnitude within participants is similar in both tasks."

4. At the end of page 14, the authors suggest that using a categorical predictor for quantifying the MC-SNARC avoids systematic underestimation of the effect size and the correlation with other measures. However, as long as the quantification of the MC-SNARC relies on the unstandardized regression coefficient (b), as in most studies, switching from a continuous to a categorical predictor is unlikely to influence the effect size. The effect size is determined by the inter-individual variability of b, not by the dispersion of the data points around the model predictions. It should be clarified that the type of predictor may affect the effect size only if the MC-SNARC is quantified using the standardized regression coefficient (r).

Thank you for this comment. It stimulated us to think more carefully about our wording. What we had in mind was that the accuracy of individual slope estimates should be higher when the statistical model being used directly reflects the theoretical model. In case of the MC-SNARC, the theoretical model is a categorical rather than linear one. This reasoning follows the general principle that the statistical model should be as close as possible to theoretical model being tested (Westermann & Hager, 2017). Upon reflection, we decided that speculating whether a better model fit of individual-level slopes (i.e., an increased signal-to-noise ratio) corresponds to a larger group-level effect size rather muddles the waters than brings sensible insights for our argument. There might be other factors involved, which we did not think of. Therefore, we have deleted the following sentences (page 17): "The use of a categorical instead of a linear function for quantifying the MC-SNARC has important consequences: First, it correctly avoids systematic underestimation of the effect size itself and an increased likelihood of false null results, and second, it avoids a systematic underestimation of the correlations between the effect and other measures like numerical or spatial skills. This means that a sometimes diagnosed "weak or non-significant" SNARC effect and its underestimated

relations to other measures might not be an attribute of the MC task but rather a result of an incorrect statistical data analysis choice."

In all other places in the manuscript, we also refrain from the statement that the strength of the MC-SNARC is underestimated when applying a linear instead of categorical model.

After discussing two theoretical reasons for the MC-SNARC to be categorical instead of linear, we added a short summary of these arguments along with a general statement that the regression predictor should be chosen accordingly (page 21):
"In summary, the rationale for a categorical instead of linear MC-SNARC is twofold: First, the intentional classification into small and large numbers is categorical in MC, and second, the interaction between the Numerical Distance Effect and the positive correlation between the SNARC effect and overall RTs contributes to a step-wise shape. Since statistical models should correspond to scientific models as closely as possible (Westermann & Hager, 2017), the MC-SNARC should therefore be tested with a categorical predictor."

- Apart from this, a critical point is missing: by increasing the fit of the model to the observed data, the use of a categorical predictor would likely increase the precision of the effect size estimate, which is highly desirable, regardless of whether it leads to an increase or a decrease in the effect size itself.

Thank you for this comment. It clearly follows our line of reasoning, which we had not expressed properly in the manuscript. We have now replaced the deleted part by the following sentence (page 17): "The use of a categorical instead of a linear function for quantifying the MC-SNARC would increase the model fit at the participant level and thereby likely also the precision of the effect size estimate at the sample level."

5. Page 19. The authors predict that the SNARC effect should be stronger in the second task due to the pre-activation of the spatial mapping of numbers. However, this prediction seems inconsistent with one of the most credited hypotheses regarding the 'categorical' shape of the MC-SNARC, which posits a direct relationship between RTs and SNARC magnitude. Since RTs are generally expected to be longer in the first task than in the second (due to a practice effect), it seems reasonable to predict that the SNARC effect should be stronger in the first task rather than the second. Both hypotheses appear logically sound; therefore, I recommend discussing and testing both (see also point 2 on page 20 and the Study Design Table). Testing both hypotheses should not change much in terms of sample size and analysis plan, since a two-sided t-test is already planned.

We thank the reviewer for the opposite argument, which we have incorporated in our manuscript (pages 22). Instead of predicting that the SNARC effect is stronger in each task when it is the second task, we now refrain from any directional hypothesis and plan to test the task-order influence in an exploratory analysis only:
"To our knowledge, only Bulut, Roth, et al. (in press; see Supplementary Materials) have tested the influence of task order, and they did not find an influence on the SNARC effect in any of the two tasks in any of three samples (130 German, 112 Turkish, and 75

Iranian participants). In fact, two opposite theoretical predictions can be made. On the one hand, the SNARC effect might be stronger in each task if it is the second, because the processing of number magnitude and its spatial mapping should be stronger if they have already been activated in a previous task. On the other hand, the SNARC effect might be weaker in each task if it is the second, because RTs typically decrease with practice and faster RTs are typically associated with a weaker SNARC effect (note that both decreasing RT and a decreasing SNARC effect over time in PJ have been found by Roth, Jordan, et al., 2024). If both mechanisms were true, they might cancel out each other and make the influence of task order invisible. Hence, we cannot make any directional prediction and will investigate the potential influence of task order in an exploratory analysis (Exploratory 1)."

Importantly, this change implied removing Hypotheses 2a (larger MC-SNARC when MC is the second task) and 2b (larger PJ-SNARC when PJ is the second task). Accordingly, we have removed the respective analysis from the section "Confirmatory data analysis" (starting on page 31) and inserted it in the section "Exploratory data analysis" instead (starting on page 37).

6. On Page 25, it is reported that participants will be asked about their level of mathematical skills. Can the author please motivate this choice, and how this information will be used in the context of the present study?

When revising the manuscript, we decided to skip this measure. The SNARC effect is claimed to be related to mathematics skills level. At the same time the effects are not clearcut with more null than positive results reported in the literature (Cipora et al., 2020 for a review, Table 1 there). Upon reflection, we decided that including this measure would not enrich our design and findings sufficiently.

In the Method, we therefore deleted the following sentence (page 29): "Participants will be asked to self-rate their math skills compared to people of their age on a visual analogue scale from very bad to very good (with responses being coded between 0 and 400 for data analysis)."

Review by Christian Seegelke, 11 Sep 2024 12:56

Summary

The main aim of the proposed experiment is to examine differences and commonalties in the SNARC effect(s) using two commonly used tasks (i.e., magnitude classification (MC) and parity judgment (PJ)). Specifically, it aims to confirm whether the typically observed different shapes of the SNARC effect between the two tasks (i.e., continuous SNARC for PJ, categorical SNARC for MJ) hold in large sample size and to assess potential correlations between the two tasks. To this end, the authors propose to conduct a large online study, already proven to be suitable to assess the SNARC effect.

Evaluation

I appreciate the study aims in assessing SNARC effects in the two most common task in a large sample size. The research questions are valid, and the proposed hypotheses seem plausible. Further, outlined analyses plans are sound, power calculations are provided, and methods are clearly and in detail described such that it allows for replication. Data handling (i.e., cleaning, outlier removal etc.) is also described and critical manipulation checks are stated, and Bayesian statistics also allow quantification of evidence in favor of the null hypothesis.

We are happy about the positive evaluation by the reviewer.

That said, my only serious concern with the current proposal is that the framing of the introduction is not tailored to the research question(s). In my view, this study is for the most part a replication attempt of the SNARC effect in two well-established tasks in a large sample while also providing a more suitable assessment of the effects (with the modelling of magnitude as either continuous or categorical predictor). However, in the introduction the authors mention many different theories (e.g., MNL, dual-route models, WM-account, polarity-correspondence), but it is my impression that the data gained from this study cannot contribute much in terms of theoretical advancement. For example, doesn't the fact that the MC and the PJ produce different SNARC shapes contradict the idea of a representation of a(continuous) mental number line? This is not to say that I see no merit in the study, I only think that the intro could be streamlined to better reflect the purpose of the study.

We agree with the reviewer and have significantly shortened the Introduction to focus on practical aspects of MC and PJ more than on theoretical considerations:

- We have cut out the entire paragraph about the cognitive-control account for the SNARC effect by Zhang et al. (2022) and only mention it briefly in parenthesis.
- We have significantly cut the "Further differences between MC and PJ" part (pages 10 and 11): "First, the MC-SNARC seems to more strongly involve visuospatial working memory, the PJ-SNARC seems to rely more on verbal working memory (Deng et al., 2017; Herrera et al., 2008; van Dijck et al., 2009). Second, the MC-SNARC and the PJ-SNARC might arise at different processing stages (Basso Moro et al., 2018; Xiang et al., 2022). Third, cognitive mechanisms underlying the MC SNARC and the PJ SNARC might differ. Namely, Prpic et al. (2016) claim that ordinality drives the SNARC effect in direct tasks (e.g., MC, where magnitude is

response-relevant), whereas cardinality underlies in indirect tasks (e.g., PJ, where magnitude is response-irrelevant). Note that Casasanto and Pitt (2019) claim that only ordinality is crucial for both direct and indirect tasks, and that Koch et al. (2023) show that order- and magnitude-related mechanisms are not mutually exclusive. Looking into these differences between the MC-SNARC and the PJ-SNARC is beyond the scope of the current study; however, the current study will provide a better understanding of the two tasks and thereby lay the groundwork for further investigations."

- We shortened the "Correlation between the MC- and PJ-SNARC" part considerably by deleting all theory-driven predictions (MNL account, dual-route model, and WM account) and focusing more on related methodological issues (pages 12 to 14): "After having described the similarities and differences of MC and PJ, the question arises whether the MC-SNARC and the PJ-SNARC are correlated. However, several properties of the two tasks might lead to a null correlation. First, the test-retest reliability of the SNARC effect has been found to be poor for both MC and PJ (correlations $.22 < r < .41$; Cipora & Göbel, 2013; Georges et al., 2013; Hedge et al., 2018; Viarouge et al., 2014). The lower the test-retest reliabilities of the MC-SNARC and PJ-SNARC, the lower is also the maximally observable correlation between the two effects. Second, the split-half reliability of the SNARC effect has been found to be poor for PJ at least in some studies (correlations $.43 < r < .96$; for an overview, see Cipora, van Dijck, et al., 2019, Table 1 there). Third, high fluctuations in the SNARC effect over time (Roth, Jordan, et al., 2024) limit the maximum correlation that can be detected. Third, at least one of the two tasks might simply not be an appropriate paradigm for assessing the underlying construct (i.e., the spatial mental number representation), making the operationalization be invalid. To conclude, low reliability, low internal consistency, and low intraindividual stability, and potentially inappropriate operationalizations of underlying mental representations might limit or cancel the measurable correlation between the MC-SNARC and the PJ-SNARC."
- We slightly shortened the "Explanations for the categorical MC-SNARC effect shape" part as well.

Finally, I have a methodological suggestion, the authors might consider. I very much appreciate the Bayesian approach to statistics, but I was wondering whether the (potential) different shapes of the SNARC effect could be evaluated more directly using a model comparison approach (as opposed to running t-tests on extracted $R^2$ values)? For example, Bayesian regression models could be specified using the brms package in R (Bürkner, 2017) and model comparison could be done using leave-one-out-cross validation with the loo package (Vethari et al. 2017). Alternatively, they could use BIC or AIC values as a model selection criterion. I just find these approaches more conventional and straight forward.

Thank you for the helpful suggestion. We have incorporated this purely Bayesian approach right after we describe that we will run Bayesian t-tests to compare $R^2$ values coming from conventionally fit frequentist models in the "Confirmatory data analysis" part of our manuscript: "Additionally, we will confirm these findings via a Bayesian approach: dRTs will be regressed on continuous and categorical magnitude for both PJ and MC in four separate Bayesian models, and in each task, a leave-one-out cross

validation will be performed to figure out which of the two predictors better fits our data (using the R package brms by Buerkner, 2017, and the R package loo by Vethari et al., 2017)."

Review by Peter Wühr, 26 Aug 2024 17:05

Summary: The authors submitted a proposal (stage 1 submission) for a highly powered online experiment addressing similarities and differences of spatial-numerical associations of response codes (SNARC) effects in two different tasks. The majority of studies has either used a magnitude classification task (MCT), in which participants classify number stimuli as smaller or larger than a reference (e.g., 5), or a parity-judgment task (PJT), in which participants classify number stimuli as odd or even, for investigating the SNARC effect. Although the SNARC effect is usually obtained in both types of tasks, they differ in processing requirements (e.g., the requirement to process number magnitude), and several differences in the observed SNARC effects have been reported. In this registered report, the authors propose a highly powered online experiment to investigate, and compare, the shape of SNARC effects in MCT and PJT, their potential correlation, and further effects of task features (e.g., task order, mapping order) on the size and shape of the SNARC effect. In the experiment, the authors intend to test a sample of 1,700 participants in standard versions of the MCT and the PJT. The experiment will have a 2 (task) x 2 (mapping/compatibility) within-subjects design. In addition, the orders of tasks and mappings will be independently varied between participants, and then used in some analyses (on task order and compatibility order effects). Several manipulation checks are planned, before the main analyses will be performed. Moreover, instead of relying on NHST, the authors will use Bayesian t tests to evaluate evidence for both the null and the alternative hypothesis.

Evaluation: The SNARC effect is among the most investigated phenomena in cognitive psychology, and has attracted researchers from many different disciplines. Nevertheless, there are still open issues concerning (a) differences in the requirements of experimental tasks that are most often used for investigating the SNARC effect, (b) the impact of basic design features (e.g., task order) on the SNARC effect, and (c) the robustness of differences in the characteristics of SNARC effects obtained with different tasks. The author‘s idea of investigating these issues in a highly powered online experiment makes perfect sense, and the results may clarify important methodological and theoretical issues. Hence, I have no doubts about the scientific validity of the research questions. Moreover, the to-be-tested hypotheses are plausible, and well justified on the basis of the literature. The authors conducted a careful and extensive power analysis for determining sample size, and they have expertly planned and described the methodologies for data collection, and data analysis. In fact, the methods are described in sufficient detail to allow for close replication of the proposed study procedures, and analysis, and to prevent undisclosed flexibility in the procedures and analysis. In summary, I cannot see a big issue that would prevent me from recommending acceptance of this proposal. Yet, I would like the authors to think about revising their description of dual-route models, and I would like to suggest additional (exploratory) analyses.

> We thank the reviewer for the positive evaluation of our submitted stage-1 registered report.

Specific comments:

(1) I did not fully understand the description of the dual-route model of the SNARC effect, and the implications for SNARC effects obtained in MCT versus PJT, as described on page 11. For example, I did not understand the statement that “number magnitude taking the fast

unconditional route in PJ should not interfere much with number parity taking the slow conditional route." In fact, number magnitude taking the fast unconditional route must interfere with the processing of parity in the conditional route (or with the outcome of this processing), because otherwise we would not observe any SNARC effects here.

In my recollection, the dual-route model of the SNARC proposed by Gevers et al. (2006) is a variant of the dual-route model proposed by Kornblum et al. (1990) for explaining spatial and other S-R compatibility effects. As correctly stated by the authors, dual-route models distinguish two parallel processing routes from stimuli to responses, a controlled (conditional) route, and an automatic (unconditional) route. Kornblum et al. applied this model to spatial compatibility effects and argued that both routes contribute to spatial S-R compatibility effects if stimulus location was relevant for the task, whereas only the automatic route contributes to spatial S-R compatibility effects (called 'Simon' effect) if stimulus location was irrelevant for the task. In particular, the spatial stimulus location will always automatically activate the spatially corresponding response, which facilitates performance when the corresponding response is the correct response, but impedes performance when the corresponding response is actually incorrect. When, however, stimulus location is relevant, a second influence on performance results from a variation of the S-R mapping between stimulus location and response location. Here, Kornblum et al. (1990) argued that processing of the compatible mapping by the controlled route is easier, or more efficient, than processing of the incompatible mapping, producing another source for the compatibility effect. Hence, when stimulus location is relevant, both the controlled route (more efficient processing of the compatible relative to the incompatible mapping) and the automatic route (automatic activation of spatially corresponding response) will both contribute to the spatial compatibility effect. This framework predicts, first, that spatial compatibility effects should be stronger when location is relevant than when location is irrelevant. Moreover, this framework also predicts (moderate?) correlations between compatibility effects in tasks with stimulus location being relevant and tasks with stimulus location being irrelevant since both tasks (or effects) have the automatic effect in common.

If we apply the dual-route logic to the SNARC effect, and to the design of the experiment proposed here, we would also assume two sources of the SNARC effect in the CMT, but only one source of the SNARC effect in the PJT. In particular, in both tasks, small numbers should automatically activate the left response, and large numbers should automatically activate the right response, producing an 'automatic' (Simon-like) SNARC effect in both tasks. In addition, one might argue that the controlled route contributes to the SNARC in the CMT, but not in the PJT. Therefore, one would have to assume that processing of the compatible (number-location) mapping is easier, or more efficient, than processing of the incompatible (number-location) mapping. Since the relevant S-R mapping and (irrelevant) S-R correspondence are perfectly correlated in the CMT, both mechanisms would contribute to SNARC effects in this task. In contrast, in the PJR, only the automatic effects of irrelevant number-location correspondence would drive the SNARC effect. There is also a variation of S-R mapping (between parity and response location) in the PJT of the present study, but this manipulation is orthogonal to the irrelevant number-location correspondence, and should therefore not (directly) affect the SNARC effect. Hence, in my view, a dual-route model would also predict (a) larger SNARC effects in CMT than in PJT, and (b) moderate correlations between SNARC effects in both tasks due to the common influence of the automatic route.

We thank the reviewer very much for the detailed elaboration and the corrections, which makes perfect sense to us. In light of the theoretical parts in the Introduction being too extensive for a replication study, as pointed out by the other reviewer Christian Seegelke, we decided to leave out theory-based predictions for a potential correlation between the MC-SNARC and the PJ-SNARC. We therefore cut out the dual-route model from the section "Correlation between the MC- and PJ-SNARC" and only mention the conditional and unconditional route very briefly later (page 19). The reviewer Peter Wühr derives very well thought-though predictions for the comparison of effect sizes in MC vs. PJ and for the potential correlation between the MC-SNARC and the PJ-SNARC, which we acknowledge in a footnote (page 19):
"As outlined by the stage-1 PCI-RR reviewer Peter Wühr, predictions can be derived from the dual-route model. Importantly, both the automatic and the intentional route are activated in MC, whereas only the automatic route is activated in PJ. Thus, the SNARC effect should be stronger in MC than in PJ because it results from both routes instead of only one route. Moreover, a positive correlation of the MC- and PJ-SNARC can be assumed based on the dual-route model, since both effects are (at least partly) caused by the automatic route."

(2) I would like to suggest additional exploratory analyses addressing the issue of differences in the shape and size of SNARC effects between CMT and PJT. These additional analyses would compare the size and shape of SNARC effects in earlier and later parts of the experiment, and possibly inform about the time course of implicit magnitude classification processes in the PJT. In the following, I will sketch some arguments why such an analysis might be interesting.

I believe that the different shape of the effects mostly reflects different task requirements, but it is possible that the shape of the effects, particularly in PJT, changes during the course of the experiment. In particular, the CMT explicitly requires participants to classify the numbers in two categories, the "small" (or "smaller than five") category and the "large" (or "larger than five") category. Hence, it does not seem surprising that this (task-dependent) classification of stimulus numbers is stronger than the task-independent processing of numerical size, and is therefore also reflected in the shape of the resulting SNARC effect. In contrast, the PJT does not require any explicit classification of stimuli according to number magnitude. Therefore, task-independent processing of numerical size – although being irrelevant for the task at hand – may occur and produce a (smaller) SNARC effect with a more linear shape than observed in the CMT. Yet, it might be possible that participants (also) begin to classify the stimuli in the stimulus set as "small" (or "smaller than five") versus "large" (or "larger than five") later in the experiment, when they have become familiar with the stimulus set. In other words, during the PJT participants might either discover that the set actually consists of two groups separated by the missing number "5", or have some natural tendency to classify stimulus sets with regard to some salient referent (e.g., the median or the modal value). Hence, it might be possible that the shape of the SNARC effect is changing from more "linear" at the beginning of the experiment, to more "categorical" at the end of the experiment. Therefore, it might be interesting to compare the shape of the SNARC effects (particularly in the PJT) between the first and second half of the experiment.

The reviewer's theory and suggested analysis are very interesting, so we added his thoughts to the Introduction (page 20):

"Note that it is possible that the PJ-SNARC is linear in the beginning of the task and becomes categorical over the course of the task, so that the continuous shape shifts to a stepwise one. That is, participants might start classifying the stimuli into the two categories "small" and "large" in PJ as soon as they become familiar with the stimulus set because they might notice that the stimulus set consists of two single-digit number sequences (i.e., 1 to 4 and 6 to 9) separated by the missing number 5. We will investigate this possibility in the exploratory analysis."

Accordingly, we added to the "Exploratory data analysis" section (page 38):
"Moreover, we will explore whether the shape of the SNARC effect differs between earlier and later phases within each task. Importantly, it is not possible to determine the SNARC effect in the first or second block of each task separately, because both blocks are needed in order to calculate the differences between left- and right-hand responses. Therefore, we will compute the models MC-4 and PJ-4 and test the resulting slopes for both the continuous and the categorial predictors against zero in two-sided Bayesian one-sample *t*-tests, but instead of considering all 30 repetitions per block, we will only consider the first or second halves of both blocks within each task (i.e., first or second 15 repetitions of each number in one and in the other response-to-key assignment). This way, we can investigate whether early trials in each response-to-key assignment lead to a different SNARC shape than late trials."

To run this exploratory analysis, we slightly adapted the procedure described in the Method section (page 29), where we originally wrote that stimulus presentation would be fully randomized:
"The order of stimulus presentation within blocks will be randomized, with the restriction that within each block, each stimulus will be presented for the 1$^{st}$ throughout 15$^{th}$ time before each stimulus will be presented for the 16$^{th}$ throughout 30$^{th}$ time (i.e., each block is divided in two subblocks indistinguishable to the participant, in which each stimulus will be presented 15 times)."

An alternative hypothesis might be that participants quite early start to inadvertently classify stimuli according to their magnitude in the PJT as well. This assumption seems plausible given the fact that SNARC effects in the PJT are mainly driven by relative numerical size (i.e., relative size in the stimulus set) rather than absolute size (e.g., Dehaene et al., 1993; Ben Nathan et al., 2009). Yet, the implicit magnitude classification in the PJT may affect the shape of the SNARC effect less strongly than the explicit magnitude classification in the CMT, leaving room for differences between numbers within the same category. If this hypothesis was correct, SNARC effects in the PJT should not occur from the very beginning of the experiment, but develop during the first blocks because participants need some time to become familiar with the stimulus set, which is the prerequisite for the implicit classification process. Hence, it might be interesting to analyze both the size and shape of the SNARC effect in the PJT as a function of experimental blocks. In contrast, one could assume that the size and shape of the SNARC in the CMT do not vary much across blocks because the explicit magnitude classification task quickly familiarizes participants with the stimulus set, and the resulting categories may dominate the shape of the SNARC from early trials on.

We thank the reviewer for elaborating on the temporal change in the SNARC effect's shape. Our newly proposed analysis, where we test the SNARC effect resulting from only the first and second halves of both blocks in each task, should provide insights (see response to the comment above).

## References


Ben Nathan, M., Shaki, S., Salti, M., & Algom, D. (2009). Numbers and space: associations and dissociations. Psychonomic bulletin & review, 16(3), 578–582.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. Journal of Experimental Psychology: General, 122(3), 371–396.

Gevers, W., Verguts, T., Reynvoet, B., Caessens, B., & Fias, W. (2006). Numbers and space: A computational model of the SNARC effect. Journal of Experimental Psychology: Human Perception and Performance, 32(1), 32–44.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: Cognitive basis for stimulus-response compatibility--A model and taxonomy. Psychological Review, 97(2), 253–270.


Signed Review (Peter Wühr)