**First round of reviews**

**Reviewer #1:**

We thank the reviewer for the constructive feedback and suggestions. We did our best to address all the mentioned points. Most importantly, we discussed the ordinality-magnitude debate, included the multiple-coding framework, and mentioned the link between magnitude and order with nature and culture in the theory section. Further, we elaborated deeper on the multiple-coding framework and clarified the theoretical impact of our work in the discussion section. As we think that we cannot address the reviewer’s comments in a brief article, we decided to change the format to a full experimental paper. Please find below our exact responses (in grey) to the comments and concerns.

General comment

The authors present an interesting study that investigates the origin of the SNARC effect. The aim of the study is to determine whether this effect is determined by the cardinal processing of number's magnitude (MNL account) or by the ordinal position of numbers encoded in working memory (WM account). To answer this important theoretical question, the authors designed a paradigm that allowed to pit the two accounts against each other without the involvement of a secondary WM task (which could trigger a WM based SNA). This aim was achieved through an innovative stimuli's selection method, which allows to discriminate if the SNAs found are determined by order/magnitude processing. Results showed that the magnitude (MNL) model explains a greater amount of variance, although the order (WM) account seems to play a significant part in SNAs as well. To conclude, the authors claim that both order and magnitude are involved in the SNARC effect.

I find that the methodology used is very clever and innovative, however crucial information is missing. What I am most concerned about are the introduction and discussion sections. In its current form, the introduction lacks information about some recent developments in the Order/Magnitude debate. As for the discussion section, I think that the results are promising but they should be better discussed to enhance their theoretical relevance. I hope the authors will find the suggestion useful and that these can help improve the quality of their work.

 Re: Thank you for a positive evaluation of our work and for bringing our attention to the ordinality vs. magnitude debate. We have now elaborated on it, which led to massive changes in both the introduction and discussion sections and helped us frame our work.

Detailed comments

Introduction

\* Please, provide line numbers to facilitate the review process.

\* I find the first paragraph to be quite vague and not providing crucial information. Considering the short word limit, authors may consider eliminating it and going straight to the definition of the SNARC effect.

Re: Thank you for those comments. We added line numbers and eliminated the first part as suggested.

\* "One effect often used to investigate": I would not say that the SNARC effect is "used" to investigate the mental representation of numbers, rather that it is an instance of the joint representation of numbers and space.

Re: Thank you. We changed this formulation.

*“One effect suggesting that numbers are mentally associated with space is the Spatial-Numerical Association of Response Codes (SNARC) effect: Relatively small/large numbers are responded to faster on the left/right side, respectively (Dehaene, Bossini, & Giraux, 1993).” (page 2, lines 21 – 23)*

\* To save words, I suggest authors to replace the terms Working Memory and Mental Number Line with their acronyms (WM, MNL) after first occurrence.

Re: As we changed the format of the manuscript to a full paper the word count is not limited anymore. As we do not need to save words and due to better readability we decided to keep the terms Working Memory and Mental Number Line.

\* "Three major approaches…most popular": I suggest going straight to the point presenting the two accounts of interest.

Re: Thank you for this consideration. We removed this part.

“*The mental number line (Dehaene, Bossini, & Giraux, 1993) and the working memory account (van Dijck & Fias, 2011) suggest different mental representations that induce this effect.*” (*page 2, lines 23 – 26)*

\* I don't think it is necessary to give a subtitle to each paragraph: I would remove the two subtitles of the MNL and WM accounts (overall the introduction is very short).

Re: Now that we expanded the MNL section, in which we also elaborated on problems with this account, we would tend to keep the section titles. As the paragraphs are not so short, we think it is more justified to have the section headers.

\* "Van Dijck et al. (2009) … mental number line account": this part is related to the WM but is in the MNL section. I think it should be removed from here. More in general, the space dedicated to the two accounts is not balanced (3 lines for the MNL and 4 + 9 for the WM). I suggest to re-organize the explanation of the two accounts in a more balanced way, linking each account to magnitude or order.

Re: Thank you for this suggestion. We have done it by linking the MNL account with the ordinality – cardinality debate and leading to the WM account. The description of the van Dijck, Gevers, and Fias (2009) findings was moved to the following section. The literature suggested by Reviewer 1 was very relevant and helpful to do it. In line with the suggestion of Reviewer 2, we have also moved the description of the multiple coding framework from the discussion to the introduction. Further, we added references, in which we also mentioned the ordinality and magnitude distinction.

\* "Which will be discussed in detail in the discussion": I suggest either to remove the entire sentence or to briefly address this issue here, to aid reader's comprehension.

Re: Thank you for this suggestion. We defined the canonical number set in the theory section and removed “Which will be discussed in detail in the discussion”.

\* The paragraph before "the current study" section looks confusing to me: "In contrast, the working memory account assumes that only the presented numbers are represented in working memory in order of their magnitude. Numbers in the range of the presented numbers, which are not presented are not represented in working memory (see examples Figure 1)". The order should be independent from the magnitude of numbers: depending on order in WM even large numbers can be associated to the left.

Re: Yes, the order can be independent of the magnitude of the numbers. However, when no special order of the numbers is primed by e.g., learning or presenting an order, Abrahamse, van Dijck, and Fias (2016) assume that the canonical number set is prompted and that this number set is ‘pruned’ to match the actually presented numbers. Hence, the order of those numbers is according to their magnitude.

“*Newer versions of the working memory account further assume that the most common numbers and their typical relations are saved in long-term memory (Abrahamse et al., 2016). One such long-term memory representation is the canonical number set representing the numbers one to nine according to their magnitude as this is the order in which those numbers typically occur (Abrahamse et al., 2016). A number item can prompt this representation and this representation is ‘pruned’ to match the presented numbers (Abrahamse et al., 2016). Hence, only the presented numbers are represented in working memory in order of their magnitude. Numbers in the range of the presented numbers, which are not presented are not represented in working memory (see examples Figure 1).*” *(page 3 - 4, Lines 56 – 68)*

\* "…their assumption how presented numbers are spatially coded": I think this sentence is not clear, please rephrase.

Re: We rephrased the sentence to “*The mental number line and the working memory account thus differ concerning their assumption about the mental representation of the numbers*”. *(page 4, lines 70 – 71)*

\* "All of them used an ordinal secondary task": I think this statement is too strong. Some studies manipulated order without asking participants to learn new ordinal sequences. Some of these used well-known numerical displays (e.g., Bächtold et al., 1998; Mingolo & al., 2021) or non-numerical magnitudes (e.g., Prpic et al., 2016) to investigate the role of order without the need to learn new sequences. The authors should acknowledge previous attempts to study order/magnitude without a secondary task. In any case, the methods used in this manuscript are different from previous literature, and still innovative and interesting, in my opinion.

Re: Thank you for this comment. To our understanding, there are two things, which need to be distinguished: the role of ordinality information for the creation of the SNARC effect, and the role of working memory sequence being actively maintained by the participants. In the study by Mingolo et al. (2021) and Prpic et al. (2016), ordinality information was recalled from long-term memory (phone keypad and musical notes respectively). In the studies we mentioned here, the role of the working memory was assessed more directly by creating an arbitrary sequence of numbers which had to be stored and recalled while performing the task. What we wanted to look into was also about using an arbitrary sequence, which was not linked to any specific pre-possessed knowledge structure (dial, notes). We are now elaborating it in the “Current study” section more explicitly. We have also added that using the indirect task (in Prpic et al., 2016 terminology), would rather promote the magnitude-related effect, so an advantage for the ordinality model would lend support to the specific WM account as proposed by van Dijck and colleagues.

\* "We expected to replicate a SNARC or an ordinal position effect in this experiment": I suggest adding "depending on which account prevails" at the end of this sentence.

Re: Thank you. We added your suggestion.

\* Overall, I think that in the introduction some not necessary information (for such a short manuscript type) is reported, while other relevant information is missing. For instance, the abovementioned studies and other recent articles on the magnitude/order debate (Casasanto & Pitt, 2019; Prpic et al., 2021) are missing, as well as the link of magnitude and order with nature and culture, respectively (Toomarian & Hubbard, 2018).

Re: We thank Reviewer 1 for suggesting this literature and helping us streamline the introduction so that we tackle the ordinality and magnitude discussion.

\* I suggest authors to carefully revise the introduction and I hope that my suggestions will help them to re-organize their arguments.

Re: We carefully revised the introduction. We thank you for the suggestions as they helped us to improve the theory section.

Methods

\* Although I think the Methods' section is well written, in the following comments I will suggest a few adjustments to help the understanding of the study.

\* Although Figure 1 caption is quite clear, I would dedicate an additional sentence to describe the stimuli's set in the methods' sections.

Re: Thank you for this suggestion. We added an additional description of the stimulus sets in the manuscript.

“*We used four number sets ([1, 2, 3, \*8\*], [2, 3, 4, \*9\*], [\*1\*, 6, 7, 8], [\*2\*, 7, 8, 9]) including one critical number (between asterisk). For the critical number, the distance toward the previous (next) number differs when regarding the magnitude of a number or its ordinal position in the sequence (see Figure 1).*” *(page 7, lines 141 – 145)*

\* I suggest clearly stating that the task used is parity judgement, and to briefly describe its instructions.

Re: Thank you. We included the description and stated that a parity judgement task was used.

“*Therefore, the current study aimed to probe the mental number line and the ordinal working memory accounts using a parity judgement (i.e., indirect) task without a secondary task (i.e., without learning a sequence explicitly and without referring to an existing ordinality-related knowledge structure).” (page 5, lines 105 – 108)*

“*Next, the participant was instructed to classify the parity of a number by either pressing the “d” or “k” key depending on the parity-to-key-mapping.*” *(page 8, lines 162 – 163)*

\* I think it is worth specifying which arithmetic task was run at the end of the experiment and why.

\* Different from the previous comment, I think description of the two final questions is not necessary and could be avoided (or at least shortened).

Re: Thank you for those comment. We specified the task. As suggested, we described the full process of the experiment including the arithmetic task for transparency reasons. However, the arithmetic task was not analyzed in the current study as this goes beyond the scope of the current research question.

We removed some of the possible response choices to shorten the description. We did not remove the questions as participants who responded that they performed the task in a very noisy or extremely noisy environment were excluded from the data analysis.

*“The experiment was concluded by collecting demographic data (cf. supplementary materials S3), running an arithmetic task consisting of 40 equations with basic arithmetic operations and a time limit of two minutes (cf. supplementary materials S4). The analysis of this task goes beyond the scope of the current research question. Then, two final questions: „How noisy was your environment?“ (“silent” to “extremely noisy”) and “If you were the experimenter, would you use the data?” (“Yes.”, “Not all of them.”, “No.”) were presented.” (page 8, lines 173 – 179)*

 Results

\* In my opinion the results' section is complete and well written. However, I suggest extending the caption of Figure 3 adding further information about the displayed data.

Re: Thank you. We added further information.

*“The four plots show the mean dRT in ms depending on either the number magnitude or the ordinal position of the numbers when they were ordered according to their magnitude. The lines indicate the averaged regression lines over all participants’ regression lines in a particular number set. For the calculation of the positions of the circles, triangles, and squares, first, the dRTs of each participant were averaged and then the mean over all participants for each number per number set was calculated. The plots in a) show the linear regression lines of all four digits of a number set. The plots in b) show the linear regression lines of the three consecutive points.” (page 12 – 13, lines 247 – 254)*

\* Always in figure 3, it looks like that the range of the y axis is not large enough. Please, change the range.

Re: We chose the range of the y-axis to enable a better distinction of the lines. But as this seems to be confusing, we adapted the range of the y-axis to show the regression lines over the whole range of the presented numbers.

Discussion/Conclusion

\* I did not fully understand the explanation the authors provide for the inconsistency due to deviance comparison. I suggest rephrasing this passage in a clearer way.

\* I think the final passage regarding the MARC effect is not much relevant to the focus of the study, therefore I would rewrite it in a more concise manner or would delete it.

Re: Thank you for those comments. We rephased the part about the inconsistency due to the deviance comparison and moved the part about the MARC effect to the supplementary materials.

\* The discussion section leads to an inconclusive final consideration. Indeed, it is not a novel finding that numbers are processed both cardinally and ordinally. This is the main problem about numerical stimuli: their order and magnitude covariate. For this reason, the order/magnitude issue in SNARC-like effects was studied using musical note values as stimuli - in which order and magnitude are clearly dissociated (Prpic et al, 2016). In the same article, it was proposed a model indicating that magnitude processing would be activated by "indirect tasks". Although authors are cautious in their interpretation, in the present manuscript they found that using an indirect task (i.e., parity judgement) magnitude prevails on order. This result is in line with the predictions by Prpic et al.'s model.

Re: Thank you for recommending this paper. We have now discussed the Prpic model in the introduction further justifying the study and mentioned it in the discussion.

\* Finally, I think the discussion of the manuscript, and consequently its theoretical impact, could be further enhanced. Authors should try to say how their data add to literature in the current order/magnitude debate (see again Casasanto & Pitt, 2019 and Prpic et al., 2021).

Re: Thank you for your input. We added this to the discussion section just before the conclusions.

Reference list (Reviewer)

 Bächtold, D., Baumüller, M., & Brugger, P. (1998). Stimulus-response compatibility in representational space. Neuropsychologia, 36(8), 731-735. [https://doi.org/10.1016/s0028-3932(98)00002-5](https://doi.org/10.1016/s0028-3932%2898%2900002-5)

 Casasanto, D., & Pitt, B. (2019). The Faulty Magnitude Detector: Why SNARC-Like Tasks Cannot Support a Generalized Magnitude System. Cognitive Science, 43(10), e12794. <https://doi.org/10.1111/cogs.12794>

 Mingolo, S., Prpic, V., Bilotta, E., Fantoni, C., Agostini, T., & Murgia, M. (2021). SNARCing with a phone: The role of order in spatial-numerical associations is revealed by context and task demands. Journal of Experimental Psychology: Human Perception and Performance, 47(10), 1365-1377. <https://doi.org/10.1037/xhp0000947>

 Prpic, V., Fumarola, A., De Tommaso, M., Luccio, R., Murgia, M., & Agostini, T. (2016). Separate mechanisms for magnitude and order processing in the spatial-numerical association of response codes (SNARC) effect: The strange case of musical note values. Journal of Experimental Psychology. Human Perception and Performance, 42(8), 1241-1251. <https://doi.org/10.1037/xhp0000217>

 Prpic, V., Mingolo, S., Agostini, T., & Murgia, M. (2021). Magnitude and order are both relevant in SNARC and SNARC-like effects: A commentary on Casasanto and Pitt (2019). Cognitive Science, 45, Articlee13006. <https://doi.org/10.1111/cogs.13006>

Toomarian, E. Y., & Hubbard, E. M. (2018). On the genesis of spatial-numerical associations: Evolutionary and cultural factors co-construct the mental number line. Neuroscience and Biobehavioral Reviews, 90, 184-199. <https://doi.org/10.1016/j.neubiorev.2018.04.010>

**Reference list**

Abrahamse, E., van Dijck, J. P., & Fias, W. (2016). How does working memory enable number-induced spatial biases? *Frontiers in Psychology, 7*. <https://doi.org/10.3389/fpsyg.2016.00977>

Mingolo, S., Prpic, V., Bilotta, E., Fantoni, C., Agostini, T., & Murgia, M. (2021). Snarcing with a phone: The role of order in spatial-numerical associations is revealed by context and task demands. *Journal of Experimental Psychology: Human Perception and Performance, 47*(10), 1365–1377. <https://doi.org/10.1037/xhp0000947>

Prpic, V., Fumarola, A., De Tommaso, M., Luccio, R., Murgia, M., & Agostini, T. (2016). Separate mechanisms for magnitude and order processing in the spatial-numerical association of response codes (SNARC) effect: The strange case of musical note values. *Journal of Experimental Psychology: Human Perception and Performance, 42*(8), 1241–1251. <https://doi.org/10.1037/xhp0000217>

van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition, 113*(2), 248–253. <https://doi.org/10.1016/j.cognition.2009.08.005>

**Reviewer #2**:

We thank the reviewer for the constructive feedback and suggestions. We did our best to address all the mentioned points. Most importantly, we revised the descriptions of the two accounts and included the multiple coding framework in the theory section. Additionally, we clarified the procedure of the analyses and conducted additional analyses as suggested by the reviewer (they can be found in the supplementary materials). In the discussion section, we stated the theoretical significance of our work more precisely and elaborated more on the multiple-coding idea. As we think that we cannot address the reviewer’s comments in a brief article, we decided to change the format to a full experimental paper. Please find below our exact responses (in grey) to the comments and concerns.

Summary:

The SNARC effect (the observation that Western subjects are faster to give left-sided responses to small numbers, and right-sided responses to large numbers), is a popular effect in the domain of numerical cognition. Several accounts have been proposed to explain this pattern of observations. The aim of the current study was to oppose two accounts: the mental number line (MNL) account (which explains the SNARC effect as a reflection of the way how numerical magnitude is mentally represented) and the working memory (WM) account (which assumed that the number-space interactions have their origin in the way how the stimulus set associated with the task is maintained in working memory). Whereas there already exist attempts to validate both accounts (which favor the WM account), the current paper takes a novel approach to test both accounts. To oppose the MNL and the WM account, different task-sets were generated each containing, besides three consecutive numbers, a deviant number in terms of its numerical magnitude (e.g., 1238, 2349, 9872, 8761). Using linear regression, it was verified whether numerical magnitude, or the ordinal position was the regressor which best fitted the data. The results were mixed. Whereas the R2 of the magnitude model provided a small but significant better fit compared to the ordinal model (42 vs 40% of explained variance), the regression based on the 3 consecutive numbers better predicted the deviant number when using the ordinal rank compared to the numerical magnitude. This pattern of results was interpreted as indicating that both numerical magnitude and serial order play a role in the SNARC effect.

 General evaluation

 The paper deals with an interesting topic in the domain of the SNARC effect. The exact mechanisms which underly this effect are indeed not yet fully understood. As mentioned in the summary, there are already other studies which opposed the MLN and the WM account (e.g., van Dijck & Fias, 2011; van Dijck et al., 2014; Huber et al., 2015; Ginsburg & Gevers, 2015), but as the authors correctly mention, all these studies used a dual task setup, which could (potentially) have over emphasized the ordinal dimension of the task. The method used in the current study is an interesting way to overcome this limitation and allows to study the SNARC effect in its "typical" context. At the same time, I was a bit surprised they only mentioned the "multiple coding framework" (which lays in the mid between the MLN and WM account) in the general discussion, despite that this is not a novel idea. In other words, there already exist data/ ideas which are in line with the conclusions drawn from the current data (Schroeder et al. 2017; van Dijck & Doricchi, 2020).

Re: Thank you for this evaluation. It is a very fair point. While restructuring the paper, we have moved the section on multiple coding (i.e., Schroeder, Nuerk, & Plewnia, 2017) to the introduction and mentioned van Dijck and Doricchi (2019) to stress that these theories have been already present before we conducted the study.

Saying this, I tend to believe that the current study lacks a sufficient level of novelty as requested for Cognition. This could maybe be overcome when (in the general discussion) a more in-depth elaboration is provided of the theoretical significance of the results. This is currently missing. Besides this, I also have a couple of other concerns which I'd like to have clarified before giving a positive recommendation for publication. These points are summarized below:

\* With respect to the hypothesis. It is true that the MNL account assumes that the SNARC effect has its origin in the magnitude of the numbers. The WM account, on the other hand, only stated that the SNARC effect emerges from the positions the numbers occupy in the WM sequence. The WM account does not specify how, and which items get into WM (at this point the WM-account is still vaguely formulated). Are this the numbers from the entire range used in the experiment (i.e., 1 to 8 or 2 to 9), or are it only numbers of the stimulus set? If it are the numbers from the range, the MNL account and the WM accounts make the same predictions. This point is briefly mentioned by the authors in the general discussion, but I think this is a central one, as it makes it a bit strange to oppose both accounts in the introduction.

Re: Thank you for this comment. We have added the information about the imprecision of the WM model, and also our interpretation of the mechanism postulated by the WM model in the theory and discussion section. As van Dijck and Fias (2011) and Abrahamse, van Dijck, and Fias (2016) use the WM model to justify the range dependency of the SNARC effect, we believed that the model assumes that only the numbers presented in the task set are being loaded. However, we absolutely agree that if it is that the whole set is loaded, the WM and magnitude effects would not be empirically distinguishable in such a setup because in such a case the ordinality would be always equal to magnitude, no matter which specific stimulus set would be chosen.

*“Abrahamse et al. (2016) explain the range dependence of the SNARC effect by assuming that when only a subset of the canonical number set is perceived this representation is ‘‘pruned’ to match the actually used items in the experiment’ (Abrahamse et al., 2016, p. 6). The WM-account does not specify, exactly which numbers are loaded into WM while conducting the task. Thus, this can be interpreted in two ways: such that only the perceived numbers are activated in working memory in order of their magnitude. In this case numbers within the range of the presented numbers that are not presented, would not be activated in working memory (see examples Figure 1). However, this wording could also be interpreted such that the mental representation includes all numbers within a particular range. However, our interpretation, is that it assumes only those numbers are activated that are actually perceived during the task. Thus, the mental number line and the working memory account differ concerning their assumption about the mental representation of numbers (see Abrahamse et al., 2016; Van Dijck, Ginsburg, Girelli, & Gevers, 2015 for detailed differences between the accounts). The mental number line account claims that numbers are represented according to their magnitude. In contrast, the working memory account assumes that only the presented numbers are represented in working memory in order of their magnitude.” (page 3, lines 61 – 75)*

\* Second, we know that the SNARC effect is sensitive to the range used in the experiment. On the one hand, you have the range dependency. This is the observation that the spatial code associated with a number, depends of the range used in the experiment (e.g., 5 is associated with right if the range goes from 1-5, but with left if the range goes from 5-9). Also, in this study different ranges are used for different participants (1-8 and 2-9). In itself this is not problematic but given the nature of the (regression) analyses, this is a weird choice, as it mixes up LTM and STM traces. Actually, the WM account was originally formulated to explain, among other observations, this range dependency. The choice for the different stimulus sets is not motivated in the paper. Are there any traces of these different ranges found in the current data?

Re: Thank you for suggesting motivating our stimulus choice. We exploratively evaluated the different ranges using the same analysis as described in the manuscript. This is reported in supplementary materials S8. The results show that for both ranges the results reported in the manuscript hold. As we do not average reaction times over participants but analyze the R2-values and the deviance calculated using the data of each participant separately, the range difference between participants should not influence our results.

To clarify our choice of number sets we described the criteria used for the number sets in the manuscript:

*“We used carefully selected stimulus sets to dissociate predictions from ordinal and magnitude number representations, always including two even and two odd numbers with three consecutive numbers and a fourth one maximally distant from the single digit range (see Figure 1 for a more detailed description of stimuli that enable this dissociation).“ (page 5, lines 110 – 113)*

\* Related to the previous point. In some previous SNARC studies only the numbers 12-89 are used to "boost" the effect. This is already something which is difficult to explain with a pure magnitude account. In the current study it is also possible that the number further way from the other gets a "very small" or "very large" status (which could give rise to a deviant dRT for these numbers). Given the context dependency of the effect, also this would be difficult to explain with a MLN account.

Re: Thank you for pointing this out. We agree that the number further away could get a “very small” or “very large” status and that there is a context dependency of the effect. Our exploratory analyses also show that there is a context dependency of the representation as the difference in the explained variance is larger for number sets where the critical value is the smallest compared to when it is the largest value. To explain these results, we propose a combined model including working memory and mental number line representations which is in line with the multiple coding framework.

\* In the analyses, the data of all stimulus sets were collapsed. I'm wondering whether this could have been the reason for the better R2 in the magnitude model. Numbers contain besides numerical magnitude also a parity status. In the stimulus sets used here, the magnitude and parity status are orthogonal in the magnitude model. When collapsing the data of the different stimulus sets, this is not the case anymore for the ordinal model. From this it is thus not surprising that the magnitude model fits the data (a bit) better.

Re: We did not collapse the reaction times of different number sets for the analyses, but only for illustrating the results. For the model comparison we calculate the *R*2-value per participant (each participant only received one number set) for both the magnitude and the ordinality model. Then we transformed the *R*2-value using the logit-function (as the *R*2-values cannot be normally distributed as the values lay between 0 and 1). The transformed values are then compared using paired t-test. Hence, the explained variance by the different models is compared for each participant.

To test whether parity differently influenced the number sets, we provide the exact results of the comparison of each number set in the additional analyses (supplementary materials S8). The results show that the difference in the explained variance between the magnitude model and the parity model is also significant when only regarding the number set 1238 for which the parity of each number is the same in the magnitude and the ordinality model. This indicates that parity cannot explain the significant difference in the explained variance between both models.

\* The effect size of the R2 is very small (only 2% difference in explained variance between both the magnitude and the ordinal model). Although I don't question the statistical significance of the difference (although I was wondering whether the raw R2 data was normally distributed; in other words, whether the transformation is necessary or not), I'm not sure whether this difference in % could be an artefact of the regression method. We know that the dRT used to calculate the SNARC effect never reflect a pure linear relationship (typically a saw-like pattern of observed). When using 1234 as independent values, all distances are of the same size, while when using 1238, this is not. Isn't it the case that because of the longer range, your goodness of fit will, in most of the cases, be better than with a shorter range (except maybe when the relation is purely linear)? Take e.g., the following dRT's. 50, -25, 25, -50 (the typical saw-like pattern) When using the 1234 as independent variables, the R2 is 50%, while with 1238, the R2 is 58%. Nevertheless, given the symmetry, you an ordinal explanation would make sense too. When you change the 8 into e.g., 200, the R2 is still 53% (which is higher than the R2 of the 1234 set, but which theoretically doesn't make sense at all). I must admit, this reasoning is based upon intuition (as I did not provide a mathematical proof of it), but it would be reassuring if the authors could comment on this (in the cover letter).

Re: As described above the R2-value lay between 0 and 1 and hence cannot be normally distributed as they are restricted to this range. This is also supported by a Shapiro-Wilk test (*W*ordinality = 0.93, *p*ordinality < .001; *W*magnitude = 0.94, *p*magnitude < .001). However, to assure the reviewer, we have also looked run a comparison of untransformed R^2. The effect remained significant (Mmag = 0.42 [untransformed R2-values], Mord = .40 [untransformed R2-values], Mdif = 0.02, t(422) = 2.65, p < .001, 95%-CIdif = [0.01, 0.04]).

We have also run a simple numerical simulation (see Koch\_et\_al\_extra\_simulation.R for the code). It shows that the *R2* associated with the magnitude predictor is not “automatically” better than the one associated with the ordinality predictor. The simulation does not produce significant differences between predictors, and the descriptive differences are in opposite directions between sets.

\* For each reported R2 two values are given in the paper, but no explanation is provided what these values mean. I suppose the first one is the average transformed R2, and the second one the one of the raw R2? Please add this information to the result section.

Re: Thank you for this hint. Yes, you are correct: The first value is the average transformed *R*2-, and the second is the average untransformed *R*2-value.

“*The magnitude model compared to the ordinality model showed significant greater R2-values
(Mmag = -0.67/.42 [mean of the transformed/untransformed R2-values], Mord = -0.96/.40 [mean of the transformed/untransformed R2-values], Mdif = 0.29, t(422) = 3.33, p < .001,
95%-CIdif = [0.12, 0.46], Cohen’s d = 0.16) and hence described significantly more variance of the data.*” *(page 11, lines 232 – 235)*

\* I like the deviant analyses. Unfortunately, the logic and motivation for this, is only minimally described. Please provide a more elaborate explanation.

Re: We have added the explanation as requested, also see the response to your comment below about the justification of the deviant analysis.

*“Additionally, we conducted an alternative analysis. We fitted a linear regression to the dRTs of the three consecutive numbers in a number set (the critical number was excluded) for each participant. We then calculated the absolute difference between the measured dRT and the predicted value from each model for the critical number for each participant (deviance) and compared them. For instance, for the set 1, 2, 3, 8, we fitted a regression line to the dRTs of the numbers 1, 2, and 3. Subsequently, we compared the dRT value for the number 8 with the value of this number being predicted by either the magnitude model (i.e., 8) or by the ordinality model (i.e., 4). The dependent measure was the deviance between the actual dRT and predicted values for both magnitude and ordinality models. This analysis was added as we wanted to look at how the regression models behave when being estimated solely based on consecutive numbers, and not being affected by the critical number (which constitutes 25% of the data), and which of these models is better in predicting the dRT for the critical number.” (page 10 – 11, lines 211 – 222)*

\* Although we originally formulated the WM account for the SNARC-effect, I also evolve to the idea multiple spatial codes are linked to numbers (see also van Dijck & Doricchi, 2019). At this point however, it is still an open question how such "multiple coding framework" can account for the different observations that gave rise to the WM account. Maybe the discussion section could benefit of an elaboration of this multiple coding idea, how it would explain data that can't be explained by the WM account, and what kind of novel predictions can arise from it.

 Re: Thank you for this comment. We elaborated more on the multiple coding idea.

Please find below some additional remarks I had when reading the paper. They are organized chronologically:

\* P2, line 15, the reference for the WM account is missing.

\* P2, line 16, is it true that the WM and MNL accounts are the most popular? I'm not sure.

Re: Thank you for those comments. We changed the formulation that the WM and the MNL are the most popular accounts and added the missing reference.

“*The mental number line (Dehaene, Bossini, & Giraux, 1993; Restle, 1970) and the working memory account (van Dijck & Fias, 2011; van Dijck, Gevers, Lafosse, Doricchi, & Fias, 2011) suggest different mental representations that induce this effect.” (page 2, lines 23 – 26)*

\* P2, Please elaborate more on the MNL and WM account, e.g., by explaining how exactly both explain the SNARC effect.

Re: We elaborated more on the MNL and the WM account and how they explain the SNARC effect.

\* P2. In principle it is not true that the effect of a visuo-spatial WM load on the SNARC cannot be explained by the MNL account. The VSWM load could interfere with the spatial aspects of the MNL.

Re: Thank you. We agree with this comment. After a massive restructuring of this section (also in line with recommendations of Reviewer 1), this sentence is not there any longer.

\* P3, line 9, please clarify "a canonical number set".

Re: Thank you for this comment. We clarified this term in the manuscript.

“*Newer versions of the working memory account further assume that the most common numbers and their typical relations are saved in long-term memory (Abrahamse et al., 2016). One such long-term memory representation is the canonical number set representing the numbers one to nine according to their magnitude as this is the order in which those numbers typically occur (Abrahamse et al., 2016).*” *(page 3, lines 56 – 60)*

\* P3, line 23, what is an ordinal fitting?

Re: As this term seems to be confusing, we changed it. We wanted to state that a secondary task such as learning a sequence could trigger the effect of the ordinal position of a number/object that was found in several experiments (Ginsburg, van Dijck, Previtali, Fias, & Gevers, 2014; Huber, Klein, Moeller, & Willmes, 2016; Lindemann, Abolafia, Pratt, & Bekkering, 2008; van Dijck & Fias, 2011).

“*One could argue that the nature of the ordinal secondary task or the memory recall triggered the effect of the ordinal position of a number/object in a sequence found in such experiments .*“ *(page 5, lines 103 – 105)*

\* P5, Design: how many participants were assigned to each between subject condition?

Re: In each condition, there were between 47 and 61 participants. The following table shows the exact distribution of participants to the conditions. When participants received a consistent parity-to-key mapping they had to react with the left key (“d”) to odd numbers and with the right key (“k”) to even numbers. In the inconsistent mapping, this mapping was reversed. We also added this table as supplementary material S1.

|  |  |  |
| --- | --- | --- |
| **parity-to-key mapping** **(first – second)** | **number set** | **number of participants** |
| consistent – inconsistent | 1238 | 49 |
| inconsistent – consistent | 1238 | 47 |
| consistent – inconsistent | 1678 | 54 |
| inconsistent – consistent | 1678 | 48 |
| consistent – inconsistent | 2349 | 58 |
| inconsistent – consistent | 2349 | 56 |
| consistent – inconsistent | 2789 | 61 |
| inconsistent - consistent | 2789 | 50 |
|  |  |  |

\* P7, did the subject knew the stimulus set in advance? And if so, how was this communicated?

Re: No, the complete number/stimulus set was never presented to the participant, because we wanted to avoid the priming of a mental representation.

“*Further, we did not present the number set used in the parity judgement task to the participants in advance, avoiding the priming of a particular number representation.” (page 5, lines 108 – 109)*

\* P8. Data-analyses. When was the logit transformation applied? Was this in the RT's, dRT's, slopes of the individual R2's?

Re: We applied the logit transformation on the explained variance (R²) of each participant to approximate a normal distribution. Slopes and dRTs were not transformed. We clarified this in the manuscript.

“*To investigate the SNARC or ordinal position effect, we tested both slopes against zero. To compare the linear regressions (model comparison) logit transformations were applied to the R²-values of the linear regressions from each participant to approximate a normal distribution. A paired-samples t-test was used to compare the transformed R²-values of the models per participant*.” *(page 10, lines 206 – 210)*

\* P8. Please provide a more elaboration for the motivation behind the alternative analyses. I also had to read this alternative analysis a few times before I understood the logic.

Re: We have elaborated on the deviance analysis by (1) providing an illustrative example of what was calculated (2) providing a further rationale for this analysis. We hope it is clear now.

\* General discussion: I found it a bit weird to read the first explanation for the limitation (the distance between the items in the ordinal model and in the magnitude model). First, although I agree that this difference might by problematic, I don't think I fully agree with the reasoning (see my point above with the small simulation). Second, this difference and probable limitation could have been anticipated beforehand, especially since this is a central manipulation in the experiment.

Re: Thank you for your critical perspective on our study design. We have provided an explanation to the point with the small simulations above. Yes, this limitation could have been anticipated before. Unfortunately, we did not. However, as we preregistered this analysis and even though it has its problems, we think, that it is important to include the analysis in our manuscript as its results and its problematic could be important for further research.

Paper reviewed by: Jean-Philippe van Dijck

**Reference list**

Abrahamse, E., van Dijck, J. P., & Fias, W. (2016). How does working memory enable number-induced spatial biases? *Frontiers in Psychology, 7*. <https://doi.org/10.3389/fpsyg.2016.00977>

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. <https://doi.org/10.1037/0096-3445.122.3.371>

Ginsburg, V., van Dijck, J.-P., Previtali, P., Fias, W., & Gevers, W. (2014). The impact of verbal working memory on number–space associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(4), 976–986. <https://doi.org/10.1037/a0036378>

Huber, S., Klein, E., Moeller, K., & Willmes, K. (2016). Spatial–numerical and ordinal positional associations coexist in parallel. *Frontiers in Psychology, 7*. <https://doi.org/10.3389/fpsyg.2016.00438>

Lindemann, O., Abolafia, J. M., Pratt, J., & Bekkering, H. (2008). Coding strategies in number space: Memory requirements influence spatial–numerical associations. *The Quarterly Journal of Experimental Psychology, 61*(4), 515–524. <https://doi.org/10.1080/17470210701728677>

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology: General, 83*(2, Pt. 1), 274–278. <https://doi.org/10.1037/h0028573>

Schroeder, P. A., Nuerk, H.-C., & Plewnia, C. (2017). Switching between multiple codes of SNARC-like associations: Two conceptual replication attempts with anodal tDCS in sham-controlled cross-over design. *Frontiers in Neuroscience, 11*. <https://doi.org/10.3389/fnins.2017.00654>

van Dijck, J.-P., & Doricchi, F. (2019). Multiple left-to-right spatial representations of number magnitudes? Evidence from left spatial neglect. *Experimental Brain Research, 237*(4), 1031–1043. <https://doi.org/10.1007/s00221-019-05483-5>

van Dijck, J.-P., & Fias, W. (2011). A working memory account for spatial–numerical associations. *Cognition, 119*(1), 114–119. <https://doi.org/10.1016/j.cognition.2010.12.013>

van Dijck, J.-P., Gevers, W., Lafosse, C., Doricchi, F., & Fias, W. (2011). Non-spatial neglect for the mental number line. *Neuropsychologia, 49*(9), 2570–2583. <https://doi.org/10.1016/j.neuropsychologia.2011.05.005>

**Second round of review**

**Reviewer #1:**

The authors did a very good job! The quality of the manuscript has much improved compared to the previous version. However, I still have some comments especially on the discussion.

The introduction is now much more complete and better represents the current status of the literature in this field. Conversely, the discussion still seems very short. Given that the manuscript is not re-submitted as brief article, I recommend authors to further discuss their results, make some speculation and recall previous literature.

Thank you for the positive evaluation of our revision.

I think the last point of my previous review was not adequately addressed:

- "Finally, I think the discussion of the manuscript, and consequently its theoretical impact, could be further enhanced. Authors should try to say how their data add to literature in the current order/magnitude debate (see again Casasanto & Pitt, 2019 and Prpic et al., 2021)."

- "Re: Thank you for your input. We added this to the discussion section just before the conclusions."

The debate on order/magnitude published in the Journal "Cognitive Science" was not mentioned in the discussion section, as well the role of nature vs culture (it was briefly mentioned in two lines of the introduction; see Toomarian et al 2018). I think there is room for some speculation that can try to integrate different perspectives, and I hope that authors can make a last effort to improve this section.

Thank you for this comment. We have elaborated on the Cognitive Science discussion now and expanded other parts of the discussion considerably. At the same time, we were unsure how the nature-nurture debate outlined by Toomarian and Hubbard (2018) informs our work. Or vice versa how we could contribute to this debate with this particular paper (as some of us have done in the past in other papers, e.g., Patro, Fischer, Nuerk & Cress, 2016; Patro, Nuerk, & Cress, 2016; Patro & Nuerk, 2017). We added one sentence on this topic to the discussion. We are happy to pick up on this important debate in greater detail if the reviewer could inform us, how our results can contribute.

**Reviewer #2:**

I've read the revised version of the paper with much attention and must say that the authors did a large effort to address the concerns of the reviewers. Although I think the paper improved a lot, I'm still in doubt about certain aspects of the paper. Below I provide a detailed description of these concerns.

We thank for acknowledging our effort and the positive evaluation of our revisions.

\* The first one is related to the novelty of the findings (I already mentioned this point in my previous review). Although the theoretical relevance of the paper is now much more emphasized and framed correctly in the introduction, the findings and conclusions are still not novel and in the current version of paper no (ground-breaking) novel theoretical insights are provided. I acknowledge that the conclusions described here are obtained with a design which comes closer to the typical SNARC paradigms (and can therefore be considered much cleaner). The results and the conclusions, however, align with what is already known. I leave it to the editor to decide whether this methodological improvement is sufficient for the standards of Cognition.

Thank you for this comment. As we state elsewhere, the results are (luckily) beyond our control, but we believe Editors of Cognitive Psychology acknowledge their value as well as theoretical implications in the light of correct theoretical framing (as emphasised by all four reviewers), worthwhile design, methodological scrutiny, and research transparency (each of these emphasised by reviewers).

\* The second doubt has to do with the analyses and the conclusions based on the results. Two series of analyses are conducted and based on the seemingly opposing findings it is concluded that the relation between numbers and space (in the SNARC effect) is driven by both magnitude and ordinal coding. In the first series, two linear regressions are conducted with the numerical magnitude or the ordinal position of the numbers in the task set as independent variables. These results show that the model with numerical magnitude as IV outperforms the other model with 2% of explained variance (42% vs 40%). Although the difference is significant, there are a couple of things that bother me to accept with sufficient certainty the conclusion of the authors that this difference is due to the magnitude coding of the numbers.

Small effect size is beyond our control. We cannot change the results. However, in the revised and considerably extended discussion we are discussing these issues in much greater detail and with critical care.

- It is a very small effect size; therefore, it makes it very sensitive to potential artefacts which have nothing to do with the coding of magnitude. E.g., you can easily replace the number 8 with a larger number, and the fit is still high. Of course, this doesn't make sense at theoretical level.

- Although the difference in R2 is significant, the effect size of this difference is very small. Given that the dRT's in parity judgement SNARC are not perfectly linear in general (in other words, there are serval other factors that have an influence on the dRT's), how safe it is to conclude that this small difference is really due to the numerical magnitude?

All these issues are now covered in the revised discussion.

- (THIS IS MY MOST IMPORTANT CONCERN) The crucial manipulation of this study is the inclusion of a "deviant number" which is in terms of numerical magnitude much smaller or larger compared to the other three numbers of the stimulus set. The results of the deviance analyses show that the dRT of this number is better predicted by the linear regression model which considers the ordinal position of the number in the task set, rather than the numerical magnitude (in this case, with a large effect size). If I understand this correctly, you can interpret this finding that the deviant number better fits the other numbers when considering its ordinal position. In the first series of regressions with all numbers are included, the magnitude model (slightly) outperforms the ordinal model. Combining both findings, you could say that this better fit is not due to the deviant number, but to a better fit with the triplet. If this is the case (that the higher R2 is due to a better fit to the three consecutive numbers), than this is not necessarily support for a better coding of numerical magnitude as for these numbers the magnitude and the ordinal position overlap. If this is true, I think this this is problematic for the conclusions drawn in the paper.

- The idea that higher r2 of the magnitude model is due to a better fit to the triplet instead of to the deviant number, makes me think that the better fit is due to a trade-off between the not perfect linear relationship between the drt's of the triplet and the numerical distance of the deviant number and its neighboring number. Afterall, the worse the fit of these dRT's with the regression line, the more the fit of the model benefits from larger gap between the final/ first value of the independent variables and the rest. This could be one of the potential "artefacts" I was talking about earlier.

We address this in the revised Discussion.

**Reviewer #3:**

This is an interesting study that investigates the role of order and magnitude in the SNARC effect. The study's aim is to contrast two theoretical accounts, i.e., the MNL account (based on number magnitude) and the working memory (based on the ordinal position). The findings revealed that the MNL account explains a greater amount of variance, however the WM account seems to play a role in SNARC effect as well. The authors conclude that both order and magnitude are involved in the SNARC effect.

The topic of the manuscript is currently debated and of great interest in the numerical cognition field. The methodology is innovative and rigorous as well as the analysis. The manuscript is well written and need only minor improvements in my opinion. I see that the manuscript has already been reviewed and the comments of previous reviewers have been addressed.

Thank you for positive evaluation of our work.

There are some citation errors in the introduction (page 2, from line 36). The authors refer to the order vs. magnitude debate citing a conference proceeding by Pitt & Casasanto (2019) instead of the (more relevant) opinion paper by Casasanto & Pitt (2019). The authors also cite a commentary to this paper highlighting an opposite view, but the correct citation is Prpic et al., 2021 (and not Prpic et al., 2016, which is a different paper referenced later in the text). The last sentence (lines 40-43) is not clearly reflecting what was presented in this series of opinion papers and should be clarified.

Thank you for spotting these errors. Indeed, we meant the opinion paper by Casasanto and Pitt (2019) and the commentary by Prpic et al. (2021). The last sentence of the paragraph was not to report the Prpic et al vs. Casasanto & Pitt discussion, but our point of view on that topic. We hope to have made it clear now after rephrasing the respective paragraph.

When discussing Prpic et al., (2016) in page 5 the authors seem to suggest that a secondary memory task was at play in the study, while this was not true. The participants in that study were expert musicians and the ordinal position of musical note values was overlearned in the same way as digits and letters of the alphabet are for everyone.

We do not suggest that. The sentence was as follows:

“(…) all of the cited studies used an ordinal secondary task (i.e., learning a sequence) or were based on some pre-existing knowledge structures. The learned sequences or knowledge structures were used as an anchor for the ordinality representation: musical notation (Prpic et al., 2016) or phone dial (Mingolo et al., 2021).”

However, to make it less conclusive, we deleted the “learned sequences” phrase, which we intended to mean a sequence learned within extensive musical education rather than on the spot during the experiment, but we can see that it may sound confusing.

The discussion of the paper is quite short and a further effort in discussing previously mentioned findings/models/accounts/debates in light of the findings of the study should be considered.

Thank you for this comment. We have now considerably expanded the discussion section.

Incorrect references:

Casasanto, D., & Pitt, B. (2019). The Faulty Magnitude Detector: Why SNARC-Like Tasks Cannot Support a Generalized Magnitude System. Cognitive Science, 43(10), e12794. <https://doi.org/10.1111/cogs.12794>

Prpic, V., Mingolo, S., Agostini, T., & Murgia, M. (2021). Magnitude and order are both relevant in SNARC and SNARC-like effects: A commentary on Casasanto and Pitt (2019). Cognitive Science, 45, Article13006. <https://doi.org/10.1111/cogs.13006>

Thank you, this has been corrected.

**Reviewer #4:**

The authors proposed an experiment to pit against each other two theories of the SNARC effect: Dehaene et al.'s (1993) mental line account versus van Dijck and Fias' (2011) working memory account. The authors asked their participants to execute what could appear as a standard SNARC experiment. However, the authors carefully selected the numbers that were used to test the participants. Instead of using random one-digit numbers as in a standard task for example, the authors used four sets of four numbers that according to the authors should generate two different predictions when considering the two above-mentioned theories. Results were not conclusive as the mental line account explains more variance but when calculating the deviance, it is the mental line account that is favored.

The manuscript is very well written, the authors (thanks also to the first round of reviews) seem to have done a very good job in order to obtain a worthy manuscript. Although the introduction is short (due to the selected type of manuscript), the authors propose a very good state of the art. The experimental paradigm is astute and elegant. The statistical analyses are sound (with a great number of participants and sufficient power) and the authors have honestly put out exactly what they have done and how. They honestly indicate that their results (and conclusions) vary according to the statistical method used, their transparence must be commended. However, I am sorry to say that 3 core elements stop me from recommending the article to be accepted in Cognition.

We thank for positive evaluation of our experimental approach and methodology.

The first problem I see, is that within the experimental set up proposed by the authors, the working memory account does not make one precise prediction but two (see around line 68). This is not problematic per se but becomes problematic when one of the predictions of the working memory account is exactly that of the mental line account. In other words, both theories make the same prediction according to one interpretation of the working memory account. From my point of view, this fact shows that the paradigm proposed is not adapted for the chosen task (i.e., distinguishing the theories).

We agree that the WM model does not make a specified prediction, however, we do not think this is a weakness of our paper (by some it can be seen as a problem with the theory though). We decided to test one, arguably the purest one of the interpretations, which is legit according to the person who laid out the theory, so we are not arguing with the strawman. Therefore, we believe that the findings may still be informative, either to refine the WM theory or to support the MNL hypothesis. Nevertheless, we extended the discussion about the second interpretation of the WM theory. We show and argue that in the normal SNARC setup (no dual task) a purely verbal account remains empirically indistinguishable from the verbal MNL account/multiple code account, as ordinalities are correlated at r = 1.0 with the number magnitudes. As long as the contributions of long-term ordinality sequences are not quantitatively specified to make quantitative predictions, almost all findings can be accommodated by both. We now discuss in detail that this might be distinguished in some future computational modelling efforts. Our experimental data provide boundary conditions for the future when such models are developed and quantitively tested against empirical data.

Secondly, as indicated above, the statistical outcome is inconclusive. Depending on the statistical analysis that is used, the conclusion seems to change. The end-result is that although the paradigm they proposed has a lot of assets, it does not allow to distinguish efficiently between the compared theories. From my point of view this inconclusiveness shows again that the paradigm proposed is not adapted for the chosen task (i.e., distinguishing the theories)

Thank you for this comment. Small effect size is something that remains outside our control, given that we ensured task reliability, sufficient sample size etc. We do not share the argument of the reviewer that inconclusive results make a paradigm inconclusive. We could have had consistent results in both analysis with ordinal or cardinal model winning in both analyses. What is more, if we left out one analysis, we could present conclusive results in one or the other direction. Since we are active in the open science and replicability movement, we tried to report the full pattern of results, even if they make our manuscript less conclusive. We do not believe that we (or any other scientist) should be punished for that.

However, we agree with the reviewer (as they also stated below) that this is a novel paradigm, which deserves to be published; we addressed the limitations raised by the reviewer in more detail in the revised manuscript, especially concerning the second analysis.

The third problem that I see, which from my point of view is the reason why the present paradigm fails to pit against each other the two theories is that the paradigm does not use the fact that both theories make clear opposing predictions when considering the order items enter working memory. For example, if we know that one participant is processing the sequence 7321 in that precise order, the predictions of the two theories become divergent, because the working memory account would predict a reverse-canonical order. Controlling order in order to create reverse-canonical order sequences in working memory cannot be achieved with the authors' new paradigm. This was the strong point of van Dijck and Fias (2011) original paradigm, although it indeed includes what the authors call a secondary task and from which they want to stay away for reasons explained in the manuscript.

While we agree that the paradigm by van Dijck and Fias (2011) is elegant and shows the role of the working memory it has specific limitations, which we discuss in the paper. Specifically, as the reviewer mentions, it required participants to memorize the sequence. As we outline in the justification of the study, this paradigm, as well as others used to demonstrate the role of ordinality somehow required the ordinality information to be processed. This on its own might have evoked the spatial mapping of ordinal information. What we were interested in was a situation when processing of ordinality is neither enforced nor even suggested by the task setup. This comes with the limitations as mentioned by the reviewer, but we wish to emphasize that the original paradigm by van Dijck and Fias does not address these research questions either.

It is a prominent discussion in the field, whether ordinal sequence SNARC effect are a natural effect or rather an artifact of the paradigm. Actually, since we are finding evidence for ordinal WM and cardinal/Long-term ordinal SNARC effects, our findings are partially extending van Dijck and Fias’ original findings, since we show that ordinal WM sequence plays an important role in the SNARC in a natural experiment at least in some situations.

Overall, I am sorry to say, that when considering the whole package brought by this article it does not, from my point of view, meet the standard for a publication in an outlet such Cognition, as the article fails to reach it designed target. Nonetheless, I believe that the article definitely deserves to be published somewhere as the paradigm is original, elegant and the topic timely.

While (for obvious reasons) we disagree with the part on publication in Cognition, we thank for the following positive words.

Line 21

The beginning of the sentence is a little bit abrupt, even more so given the fact that it is the first sentence of the article. However, given the restrictions in term of length of the article and the previous comments of the reviewers I do not insist. However, I think that non-expert readers would have difficulties in understanding what follows: ": Relatively small/large numbers are responded to faster on the left/right side, respectively". Maybe the sentence could be unpacked a little bit?

Thank you for spotting that – we have now rephrased to facilitate accessibility.

Line 54

I cannot understand "This association is created by binding a number items". Or maybe there is just a typo.

It was a typo indeed, it should be “(…) by binding number items”.

Line 66

The authors write within brackets: "see examples Figure 1". I don't think it is clear to the reader how Figure 1 helps the comprehension. Could the authors be a little bit more explicit?

This has been now clarified with references to specific parts of Figure 1.

Line 69

I am little bit troubled with the sentence: "However, our interpretation, is that it assumes only those numbers are activated that are actually perceived during the task." It makes the authors look a little bit cavalier.

See response above. The WM account remains unclear according to J.-P. van Dijck, who first proposed and developed it. In our view it seems open space for different interpretations, and in the manuscript, we are discussing both of them, at the same time are able to empirically test one of them. We have rephrased the sentence to make it sound more neutral.

Line 166

I was wondering why 75%?

We acknowledge that this threshold is arbitrary. We found this one to be reasonable to ensure the participants understood the instructions and can proceed from practice to experimental session. When setting up this threshold we considered that (1) the chance level in this task was 50%; (2) that this is a relatively simple task as knowledge of the parity of single digit numbers is common among educated adults in Western cultures. Even though thresholds of e.g., 70%, 80% or even 73% would be equally good and meet the criteria we outlined, we had to make some decision and decided for 75%.

Line 196

I know that Cipora et al. (see also van Dijck & Fias, 2011) used that cut-off but nonetheless what is the rationale for 250ms?

We acknowledge that this is an arbitrary border, but at the same time we find it quite reasonable: correct choice reaction times shorter than 250 ms are physiologically unlikely. Again, one might argue that the threshold of 200 ms (0.49% of trials excluded below that threshold) or 300 ms (1.19% of trials below that threshold) would be equally good / better but to the best of our knowledge, there is no set standard to this, and we had to make some decisions about treatment of anticipations. Anyhow, the results do not strongly change with different boundaries (model comparison, 200 ms: *t*(422) = 3.22; *p* = .001; [0.11, 0.45]; 300 ms: *t*(414) = 3.42; *p* < .001; [0.12, 0.43]; deviance comparison, 200 ms: *t*(422) = 15.58; *p* < .001; [48.45, 62.44]; 300 ms: *t*(414) = 15.49; *p* < .001; [48.49, 62.58]). Please note that with this threshold much fewer than 1% (i.e., 0.59%) of reaction times was classified as anticipations and discarded. The raw trial-level data along with our analysis script is openly available on the associated OSF profile and interested readers / methodologists are free to evaluate other thresholds.

Line 200

I was wondering why 75%?

As in responses to above points, this 70% threshold is arbitrary (not 75% as pointed by the reviewer). We found this number reasonable to be able to conclude that the available reaction times for this number / response side configuration originated from a meaningful response to the given number. We wanted to avoid the situation that a participant responds at the chance level to a specific number with a given hand but the reaction times corresponding to correct responses (even though probably not very meaningful) are still included into the analysis. Please note that the fewer reaction times included the less reliable the corresponding mean would be.

As regards the justification of all these choices please note that all these were determined before we collected the data, as outlined in the preregistration. Even if arbitrary, these decisions had to be made and they were not conditioned on our data. At the same time, using the shared trial-level data and analysis script, interested readers and methodologists can check whether and how different thresholds would affect the results.

Line 276

As stated above, it is problematic that through this paradigm one interpretation of "the working memory account would make the same prediction as a multiple coding", as it makes this whole endeavor less valuable.

As outlined above, we are unsure whether this is necessarily a problem with our paradigm, or some sort of lack of clarity of the WM account. In particular, one version of it seems to be empirically indistinguishable from the magnitude/multiple coding account in case of single digit numbers. We have now elaborated on this issue in greater detail in the revised discussion.