Response to Revision Round 2

"Both reviewers are largely happy with your revisions, but would like some minor revisions mainly to tie down remaining analytic flexibility, to do with the use of directional vs non-directional tests and also outlier exclusion."

We are grateful for the editor's and reviewers' thoughtful comments and suggestions. We feel that the RR has further improved as a result. Below please find our responses to each comment. Respective changes have been highlighted in the manuscript. We hope these revisions meet with your approval.

Before we move to addressing each comment, we would like to mention that we made a further addition to the revised manuscript. While working on this revision, we identified two more studies that used the dual-task paradigm to determine the role of WM in adults' symbolic number processing (Herrera et al., 2008; van Dijk, Gevers & Fias, 2009). Although those studies addressed very different research questions regarding adults' symbolic magnitude comparison performance focusing on the SNARC effect, we found them informative and relevant to our study because they demonstrate that the VSSP may play a role when processing the spatial representation of number. Thus, we have included these studies in the introductory section of our revised manuscript (pp. 13-14).

Responses to comments:

1) I would also like some more scientific justification for the minimal effect sizes chosen. You refer to past papers for deriving them, but I wasn't sure how you infer a minimally interesting and plausible effect from those papers. For example, why 100 ms? Would an effect of 50 ms really not be interesting (or else not plausible)? Why would 5 trials be just interesting? I realize these are nice clean numbers, and hence simple, and that is something going for them; but I worry a non-significant result might not mean much if the true effect was e.g. 50 ms and this would have been very interesting scientifically if known.

Thank you for raising this point. Prevailing theories of number processing such as the ANS and Triple Code model have been developed for explaining individual differences in accuracy. Therefore, we have now based our power analysis on the minimum effect size of interest in accuracy. Specifically, we calculated our minimum effect size of interest by considering the smallest relevant decrease in performance. Based on estimates of adult performance on standalone comparison tasks (e.g., dot comparison accuracy: 99.7% (SD = 0.3) (Lyons et al. 2012), we powered to detect a difference in 5 out of 160 trials on the primary task, which would reflect a 3% difference. Based on these calculations the largest required sample size was N = 81, and therefore this is the sample size we will recruit for this experiment. For RT, this would allow us to detect differences of 50ms for the symbolic and nonsymbolic comparison conditions and a difference of 80ms for the cross-modal comparison condition.

We outline this reasoning in the revised manuscript in pp. 17-18.

2) I'm mostly happy with the revision, however, I do have a few points to note. In the table in Appendix C some of the hypotheses that have been stated are explicitly directional and

some are not explicitly directional when they should be. For example, RQ3b: "This means we will expect to see a decrease in performance in the non-symbolic primary task in the VSSP dual-task condition in compared to the standalone condition for large quantities, but no decrease in performance for small quantities."But then "If t-tests for both small and large quantities are significant ($p < .05$)" Surely this should be directional. or in RQ1b there's a mix of "If t-test is significant at $p < 0.05$, and indicates that performance is lower" (which is directional) but then "but the t-test indicates that there is no significant difference between the VSSP dual-task" (which is non-directional). If the hypotheses are directional then they should be clearly stated as directional. Furthermore, since two-tailed tests are being used, it should be clarified whether "non-significant" means two-tailed non-significant or whether it just means a result NOT "significant in the specified direction" (that is, one-tailed non-significant: i.e., $t(10) = 2$ gives a two-tailed p of ~.07, but a one-tailed p of either ~0.036 or ~0.96 depending on the tail that one looks at). For example, what if there is a significant difference, but just not in the predicted direction. For this reason, I think it might be easier to use one-tailed tests (at an alpha of 0.025) rather than two-tailed tests (at an alpha of 0.05) when the direction actually matters and only use two tailed tests when the direction doesn't matter.

Thank you. Although for some RQs we had had directional hypotheses comparing dual to stand-alone performance, we do not have clear directional hypotheses for comparing the primary task done under PL vs VSSP dual task conditions. We had always intended to do two-tailed tests; therefore, in the revised manuscript we now only mention non-directional hypotheses (see Appendix C, pp. 36-44).

3) Thanks a lot for implementing our comments so rigorously, especially specifying the analysis decision tree.I have only one more comment on data pre-processing regarding the following sentence in the introduction of Appendix C: “Outliers will be examined for each condition (all combinations of primary and secondary task) and extreme outliers (> 3.29 SD, Field, 2016) will be removed from the analysis for that condition.”As it makes a difference whether you look at group means or individual means per condition, I would like you to clarify this here. Furthermore, I do see it critical to look at outliers per condition, as Andre (2021) outlined in a recent simulation study that data pre-processing needs to be blind to the research hypothesis and addressing outliers separately per condition can artificially induce effects. André, Q. (2021). Outlier exclusion procedures must be blind to the researcher’s hypothesis. Journal of Experimental Psychology: General.

Thanks for your thoughtful comments and helpful suggestion. We have now revised our outlier exclusion procedure accordingly. Specifically, outliers will be examined for performance on each task (i.e., primary and secondary tasks). Extreme outliers (> 3.29 SD, Field, 2016) will be removed from the analysis (Appendix C, pp. 36)