Editor

Sorry for the delay in getting back to you about your submisison, I sent out 33 review requests - and I have now obtained reviews from 2 experts. Both have very helpful comments to make about improving the manuscript. Loenneker requests some clarifications regarding the introduction and methods. Note there is no need to write an anticipated discussion at this point. Colling makes the point "a power analysis is only as valuable as the effect size estimates that go into it"; that is, for power, and equivalence tests, the point is to find the minimal effect of interest, as justified by your particular scientific context: See here for advice https://psyarxiv.com/yc7s5/. A minimal effect of interest is different from a roughly expected effect, but you use both terms; further, you do not argue why the effects you chose are minimally interesting ones. Colling suggests the use of Bayes factors; these do incorporate roughly expected effect sizes. (See previous reference.)

Thank you for your significant work obtaining reviewers.

There is no previous empirical work that allows us to make specific predictions of the size of the effects. Therefore, our power analysis was based on minimal effects of interest. We have now clarified this in the manuscript (p. 16).

For accuracy, there are no previous studies that have reported accuracy on the primary or secondary task for a magnitude comparison task performed under dual-task conditions. We believe a difference of 5 trials (out of 160) on the primary task and 2 trials (out of 20) on the secondary task would be the smallest effect to be meaningful, particularly given that these are tasks with typically high accuracy when performed in isolation. Therefore, we used these as the basis of our power analysis (based on the means and SDs reported by Lyons et al (2012) for adults performing the magnitude comparison task in isolation). Typical effects sizes for accuracy on dual task studies where the secondary and primary task draw on the same working memory components are much larger (e.g., Cragg et al., (2017) found an impact of dual task conditions on arithmetic of d=0.4). Therefore, we are confident that the impact of our dual task manipulation would be detectable if working memory is recruited to perform these tasks.

For RT, we selected a difference of 100 msec to be the smallest effect to be meaningful. Here, there is one relevant study to inform our estimates. Maloney et al (2019) reported RT for adults performing a dual-task non-symbolic comparison study. The impact of the dual task condition was 150msec. Therefore, we are powered to detect a smaller effect. We believe that an impact of less than 100msec (with expected standalone RT of approx. 800msec) would not be big enough to be theoretically meaningful.

One further comment both on Colling's advice and on your planned analyses: The aim of a Registered Report is to tie down analytic flexibility. Thus one should stick to one inferential approach that allows justification for asserting no effect or an effect. That could be power, equivalence testing or Bayes factors. Stick with one in the pre-registration. Note your procedure of performing a significance test against 0 and then following up with an equivalence test if the former is non-significant can lead to a inconsistency because different models are used for the two classes of test: A t-test against 0 can be significant and yet an equivalence test would have shown equivalence had it been done. So the way to use equivalence testing as one coherent procedure is to generalize it as "inference over intervals": Determine if the X% CI lies inside or outside the equivalence region (assert/reject equivalnce respectively) or straddles it (suspend judgment). Colling's point remains for this as well as power: The procedure only makes scientific sense if the equivalence region is scientifically justified as the region of no interest.

Having said all that, the reviewers agree the design is well thought out to tackle the issues you wish to address, so I very much look forward to seeing a tidied up manuscript.
best
Zoltan

Thank you for these helpful comments. Following your suggestion, we now focus only on the power-based comparisons as outlined in the analysis table. The equivalence tests were added to be used only in the case that the t-tests comparing standalone conditions and dual-task conditions were non-significant. As you have highlighted this could have resulted in undetermined outcomes if the results were inconsistent. We have therefore removed reference to these in the analysis plan.

Reviewer 1: Reviewed by Hannah Dorothea Loenneker
The authors conceptualized a thorough study, investigating whether adults can automatically process and translate between numerical representations. Their research can be an interesting foundation for further studies, as they're aiming at deepening the understanding of the link between numerical entities coded in different modalities within human cognition. They use the experimental design of a dual-task set-up, with different secondary tasks to assess whether working memory is a necessary prerequisite for (non-)symbolic magnitude comparisons and cross-modal translations.

R1.1 I would like to acknowledge the authors' intention to make their raw data publicly available and would like to stress the importance of well-documented meta-data with accessible descriptions of the respective variables (following the FAIR criteria).
I feel like there are some small language errors in your manuscript – please revise.

Thank you for these positive comments. We have carefully proof-read the manuscript to check and correct minor errors.

R1.2 Page 5, last paragraph: I wouldn't agree that the first numerical representation children acquire is the verbal one, as studies implementing the looking time paradigm already show an increase in precision of the ANS between the ages of 6 to 10 months (e.g., Brannon, Suanda, Libertus, 2007; Lipton, Spelke, 2003, Wood, Spelke, 2005; Xu, Spelke, Goddard, 2005).

We apologise that we were unclear. The verbal representation is the first exact symbolic representation that children acquire. We have added this to the text (p. 4).

R1.3 Page 5: "The ANS is assumed to provide estimates of the numerosity of a given set. Repeated presentations of the same numerosity result in varying points of activation and consequently representations of quantity in the ANS are approximate." -> it's not quite clear to me what you're trying to say here. Can you reformulate?

We have rephrased this and refer to Gallistel and Gelman's (2000) paper which further explains the continuous accumulation process for interested readers. We hope this is now clearer (p. 5-6).

R1.4 I would probably introduce the role of inhibition and visuo-spatial skills for performance in ANS tasks when describing the association of ANS and arithmetic on page 6. This already gives a hint for your later research question that domain-general cognitive abilities may play a role for the associations between the different representations.

Thank you. We have now added a mention to this on p. 6.

R1.5 Can you clarify why working memory is the only domain-general cognitive function you expect to modulate the relationship between the different numerical representations? Considering the literature in this field, I do see several other candidates such as e.g., inhibition, verbal skill, or visuo-spatial skill.

We agree that other domain general skills may potentially be involved in the processing of different numerical representations, such as inhibition. This is now mentioned on p. 6 (see also response to R1.4).

The reason we selected working memory is because of the focus of our study, namely the question of whether we process and translate numerical representations automatically or not. As explained on p. 11, there is no WM involvement in tasks that are automatised (Ding et al., 2017). Conversely, if there is, one can conclude that the task at hand is not processed automatically. In the present study, we adopt the experimental dual-task paradigm, which is used to unravel the role of WM components in processing a given cognitive task. We aim to examine participants' performance on our three primary tasks (dot comparison, digit comparison and cross-modal comparison) under both phonological and visuospatial interference conditions, therefore our findings will inform us about the involvement of verbal and visuospatial skills. In our discussion section, we will suggest that the role of inhibition is addressed in future studies using alternative appropriate methods.

R1.6 What is your rationale for such a broad age range? With working memory and executive functions decreasing with age, I would suspect that you might find differences between age groups. Will you correct for these (for example adding age as a covariate in case it is significantly correlated with your outcome measure)?

Our selected age range is 18-65, we apologise for not clarifying the reasoning behind this choice. Studies examining working memory capacities across the lifespan have shown that there is relatively little change in working memory performance in adulthood (Alloway & Alloway, 2013). In particular, Alloway and Alloway (2013) showed that people in their 60s perform at a similar level to those in their 20s; stark declines in people's visuospatial skills were observed over the age of 70. Thus, we included people up to 65 years of age. We now mention this rationale in p. 16.

R1.7 Thanks for sharing your sample size considerations so transparently in the Supplementary. I'm just not quite sure how you come to the conclusion that expected decrease in performance in dual-task condition (number of trials difference or increase in RT) is 5 trials or 2 sequences. Could you elaborate on that a bit more in the manuscript?

Please see response to the editor above. We have now clarified that these are smallest effect sizes of interest, rather than expected effect sizes.

R1.8 Will you apply some kind of manipulation check that participants actually tried to stick to the secondary task? One could imagine some participants to only focus on the primary task resulting in low performance in the secondary task but a low effect on working memory load as well. You state that you will record accuracy of recall for both secondary tasks, but are you planning to control for this by excluding participants not reaching a certain level of performance regarding the secondary task?

We agree that it is likely that participants may focus on the primary task and neglect the secondary task. In this case we may not observe a dual-task effect on the primary task. However, we would expect to see lower performance on the secondary task compared to completing this in

a standalone condition. Therefore, we will analyse accuracy on both the primary and secondary task, as described in our analysis plan.

R1.9 Do I understand your design correctly: one participant will have both secondary tasks alongside each primary task so that the effect of the respective secondary tasks can be compared within participants?

Yes, this is correct.

R1.10 In my view, the section on planned analyses lacks specification: Could you please lay out a decision tree, starting with testing the assumptions of your planned statistical analysis and then elaborating which changes you will make if the assumptions are not met (e.g., data transformation, robust hypothesis tests, etc.). Generally, I miss a section on your planned pre-processing pipelines, as these can heavily influence your results and should therefore be pre-registered as well. Will you only consider reaction times or accuracies as well?
Please consider possible limitations like the fact that you're only using single-digit material which doesn't allow for generalizations to multi-digit material. What if your material is too easy, resulting in ceiling effects and low variance?

We apologise that our analysis plan wasn't sufficiently detailed. We have added an initial introductory section outlining the preliminary analyses and data checks to be performed (p. 35). As described in the table we will be analysing both accuracy and RT on the primary task, and accuracy only on the secondary task. We will conduct all the analyses listed in the table (we have now removed the equivalence tests which were previously dependent on the outcomes of earlier analyses).

We expect accuracy to be high on the standalone versions of the task (hence analysing RT as well). Under dual-task conditions impact may therefore be detected on accuracy or RT. Using single-digit material does contribute to these high accuracies, however, this is essential for us to answer our theoretical questions (e.g. based on differences between and within the subitising range).

R1.11 As I believe the PCI design table to be very helpful in thinking through possible outcomes, it would be great if you included these considerations not only in the Supplementary but in an anticipated discussion section as well. One point regarding the design template: I think it would be more straightforward if you enumerated the three RQ3 as RQ3a, RQ3b and RQ3c as I was a bit surprised by the repetition of RQ3 which is only explained by the differences in the hypothesis column.
I'm looking forward to seeing the results of your interesting study.

Thank you for this suggestion. We have renamed the hypotheses as suggested, and slightly re-worded them to make the differences between RQ3a, RQ3b and RQ3c clearer (p. 41-42).

Following the editor's guidance we haven't added an anticipated discussion section.

Reviewer 2: *Reviewed by Lincoln Colling*

R2.1 The logic of the proposed experiments are clear and the interpretation of the results with reference outlined theory is also clear. I would just make one note about the phrasing on one point in the introduction.

In the introduction it's said that we represent numerical information "with Arabic symbols". Surely the representations refer to/are about Arabic symbols. The current phrasing seems to imply that the representational vehicle itself are Arabic symbols, which doesn't seem correct.

Thank you for your positive comments. We have edited the abstract and early part of the introduction to clarify that we are here referring to internal representations of number: verbal number words, visual Arabic numeral form, non-symbolic (analogue magnitudes). We have cited the Triple Code Model to indicate that we are referring to these codes. The Triple Code Model is described in more detail later in the introduction.

R2.2 The methodology also appears to be sound. However, I do have a few comments on the power analysis because I found it difficult to exactly follow in some places.

There is nothing wrong with the power analysis per se; however, a power analysis is only as valuable as the effect size estimates that go into it. The crucial power analysis is the analysis based on the findings of Lyons et al (2019), which give a sample size of 81 based on a *d* of 0.33. However, I don't feel that there is really sufficient justification given for this effect size. It would be valuable to include a little more information about how this exact value was obtained and the reasoning behind it. This could be as simple as also including the variance estimates together with the raw effect estimates (e.g., "an increase in primary RT of 100 ms (SD: xxxx)"

Please see response to the editor above. We have now clarified that these are smallest effect sizes of interest, rather than expected effect sizes. We converted our smallest effect of interest (in terms of the number of trials) to effect sizes based on the estimates of standalone performance taken from Lyons). We have added the details to the manuscript as requested.

R2.3 Furthermore, bearing in mind that power analyses are only as good as the effect size estimate that goes in the them, and bearing in mind that coming up with an effect size estimate is fraught with many difficulties, and furthermore, bearing in mind that null hypotheses significance tests can be difficult to interpret in the absence of significance, I feel it might also be valuable to supplement the analysis plans with some additional simple Bayesian tests for the paired analyses (that is, to supplement the proposed *t* tests). I note that while the authors do also propose equivalence tests, not enough information is provided about what would constitute the bounds of equivalence.

Following the editor's suggestions, we have not added Bayesian analyses to our pre-registration and have removed reference to the equivalence tests.

R2.4 In addition to the power analysis, I also thought there was many something missing with regards to exclusion criteria for trials or participants.

I wasn't able to find any information about whether participants would be excluded for low accuracy or long reaction times. This would be useful to know.

We apologise that this information was missing. We have now added further information in Appendix C regarding our pre-processing. We will analyse median RT to reduce the influence on long reaction times. For both accuracy and RT we will examine outliers for each condition (every

combination of primary and secondary task). Extreme outliers (> 3.29 SD, Field, 2016) will be excluded from the analysis of that condition only.

R2.5 Further on this point, for the non-symbolic task, will congruent and incongruent trials be analysed separately at all? Setting a minimum performance level for incongurent trials might be a way to ensure good data quality.

Our pre-registered analyses do not include plans to analyse congruent and incongruent trials separately as it is not essential for our primary research questions and congruent and incongruent trials are intermixed within the primary task blocks.

Most investigations of the non-symbolic comparison task consider overall trial performance and it is difficult to determine exactly the processes involved in the congruent and incongruent trials. Our previous work has found that there are wide individual differences in the extent to which participants are influenced by congruency. We are interested in the involvement of working memory on the translation between symbolic and non-symbolic quantities, regardless of the processes involved (inhibition or ANS) and therefore we wouldn't want to exclude participants simply on the basis of low performance on incongruent trials. However, we agree that this may be something of interest to consider in exploratory analyses.