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.

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)”

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.

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. 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.