Dear Prof. Chambers,

We are glad that you found our manuscript already very close to meeting the Stage 1 criteria, and we thank you for having gathered the opinions of three experts in such a short time. Similarly, we would like to thank the reviewers for raising a number of extremely good points – we've done our best in order to address them below. All changes (addition/deletions) on the manuscript are highlighted in red, and we point you and the reviewers to the exact pages and lines where you can find them.

We found that reviewers were all in agreement on many points. For instance, they judged our exclusion criterion based on floor and ceiling effects at training to be too harsh and to be detrimental to our goals. After careful consideration, we have now come to the conclusion that they were right and therefore we have dropped it, however we retain exclusion criteria based on the vocabulary tests based on results of Pilot 1 (pag 48, lines 1295-1300).

The second point that has been touched by all reviewers was related to how we computed our chance levels in the pilot and in the main experiment. We have realised that we didn't explain with sufficient detail how chance level was worked out based on the possible moves. We have now clarified which moves were possible in pilot 1 and which are possible in the proposed experiment by adding two dedicated sections and additional figures explaining how chance level is computed (Proposed methods: pag 38-39, lines 1009-1029, figure 8; Pilot 1: pag 65-66, figure 13).

In other instances, we realised that concerns were caused by not having reported crucial information. This is the case of the anonymous reviewer's concern regarding performance in the noun training tasks, and Prof. Rowland's issue regarding the lack of details for measures of individual variabilities. We now provide detailed info about the noun training tasks (pag 35, lines 974-986) and participants' performance in the results section of pilot 1 (pag 66-67, lines 1721-1725, table 7). Finally, we have removed all references to individual variability tests, which we have realised are not feasible. However, we have added details about a pre- and post-test memory span task (pag 41, lines 1063-1070). We do not set up predictions for this task as we approach it in an exploratory manner.

We have also addressed Prof. Rowland's concern about our study using a high prior – we note that in terms of computing Bayes Factors, our choice to use a prior that stems directly from our pilot aims to avoid introducing a bias towards either H1 or H0. Moreover, if our prior is optimistic (as suggested by Prof. Rowland), we are setting a very high bar for accepting H1, making our approach conservative. However, there is no reason to think that data collected online with this paradigm will be noisier than in person. By comparing the variability of online data (in the former pilot 2) and in person (pilot 1) we see that the SDs are in the same ball park (see point 18 in this letter). This suggests that there is little risk of data collected online being noisier than in person, at least in our experiment whereby we compute accuracy and not RTs, which is more influenced by in-lab vs in-person differences.

Finally, we have addressed Dr Mayor's point about the fact that we did not test word order, despite this being a critical aspect of the constructions being learned. On consideration, we decided that this was an important point and we have therefore made changes to our proposed methods to reflect this. In particular, we have proposed a different grid configuration whereby we present both nouns heard in the sentence among the movable and fixed objects, along with distractors (pag 36, lines 995-1001. Pag 37, Figure 6). This configuration allows us to disentangle whether children have learned word order as well as the meaning of above/below. We provide a schematic representation of these moves in figure 8, pag 39, and a detailed plan of analyses involving these moves in pag 54 (from line 1473) to pag 56.

As we discussed with you by email, we feel that these changes (in combination with some others also made in response to reviews and detailed below) are sufficiently substantial to warrant a new pilot, which will replace “pilot 2” from the previous submission, which is no longer directly relevant. Following your advice, we have not yet run this new pilot, but we have provided a new link to the reviewers to try the new design themselves (this has been also updated at pag 74, point 4 – Data and materials availability section). We look forward to receiving yours and the reviewers’ feedback on these proposed changes. Note that the results section for pilot 2 is currently empty, to be completed once we have that data. However, we provide in this section a short paragraph where we list what we are going to report once we get data in (pag 72-73, lines 1831-1870).

We hope that you and the reviewers will find the new version of the manuscript compatible with collecting new data for pilot 2. We have genuinely appreciated this opportunity to improve our design provided by PCI. At the same time, there are places where we have stood by our decisions and we have tried our best to explain our rationale as clearly as possible below.

Thank you again for the time you have devoted to this manuscript.

Best wishes

Eva (on behalf of all the authors)

Reviews

Reviewed by anonymous reviewer, 27 Jan 2022 00:30

1. *It has been a great pleasure to read the manuscript: it is very clear, exceptionally-well structured and thorough. The authors performed computational simulations, implementing an error-driven, discriminative learning process, and two pilots to ensure the validity of the design. The research questions make sense in light of the presented theories and they are clearly defined. The protocol is sufficiently detailed to enable replication (see some minor suggestions though). The hypotheses are clear and the authors have explained precisely which outcomes will confirm their predictions. The sample size is justified. The authors need to provide data quality checks, though. The authors clearly distinguished work that has already been done (e.g. Pilot 1 and 2) from work yet to be done. Although I am overall very impressed by the quality of the manuscript, I would like the authors to address the below-mentioned points, before I can recommend in principle acceptance.*

Thank you for your kind words- we are glad that the you had a general positive impression of our manuscript.

2. *I thank the authors for providing a link to the experiment. It has been a lot of fun to play it (the whole family enjoyed). However, after having passed the experiment myself, it seemed to me that, in the high variability condition, learners are in a more disadvantaged position, as compared to the other two conditions: In addition to learning the adpositions, the high-variability group has a concurrent task to perform - to recognize and learn Japanese words in running speech (to adapt them and connect with their English equivalents) across different sentential positions and noun combinations, which might hinder learning of the adpositions per se to a larger extent as compared to the low-variability group (cf below).*

There are a couple of points about our methods which we think are relevant here (we also acknowledge that we were not fully clear about these in the previous version of the manuscript).

First, stimuli were not actually recorded as whole strings but instead in two parts which were then concatenated. The first was the first noun + object marker (e.g. “banana o”) and the second the location noun + post-position + verb (“oku”) (e.g. “chocolate noueni oku”) (note that this was not clearly explained in the previous manuscript and this has now been clarified, i.e., page 35-36, lines 974-986). The main reason for recording in this way was practical: it allowed us to reuse the same recording of (e.g.) “banana o” across different sentences, reducing the number of stimuli that needed to be recorded (especially since we randomized assignments of nouns across participants). While this has the disadvantage of making the stimuli less naturalistic, it does have the advantage of reducing the difficulty in terms of adapting to the nouns in context, which may go some way to allay the reviewers concerns about the greater difficulties that children might have in the high variability group.

A second point is that, contrarily to the demo (whose goal was to show only the learning phase), by the time they get to that phase children in both conditions have already been familiarised with the nouns in their training set in the first “noun practice” stage. There is also a second “noun practice” test just before testing using the nouns for that child test set. In each case, sessions involve hearing a

noun and identifying the picture that goes with that noun. We have now added in the data from these into our pilot (pag 66-67, lines 1722-1725, table 7) – they show ceiling performance in both conditions (see also our response to your later comment – point 4 in this letter).

We believe that the above makes it somewhat unlikely that any differences between conditions are due to difficulties identifying the nouns in context. To further mitigate against this possibility we intend to use low performance in these noun practice tests as our inclusion criteria (pag 48, lines 1295-1300; see response to your point 4 below).

Finally, for our main analyses of training and testing data, we will be further excluding all trials where children did not correctly identify the two nouns involved in the sentence (We make clear this in the "exclusion criteria at the trial level", pag 48, lines 1271-1282). This means that if we do see differences between conditions in these main analyses, they cannot be driven by differences in ability to identify the nouns. On the other hand, we will include additional analyses which will check whether there are more such excluded responses in one condition than the other (see our response to your point 9 below), as in our sanity check analysis at pag 49-50, lines 1310-1342.

3. *Accent adaption can be challenging and cognitively demanding for children, using the resources necessary for learning the target adpositions. On the other hand, this can provide an advantage when exposed to novel speakers and novel items (as tested in the current study), as the high-variability group would have had learnt to adapt Japanese accents more efficiently, which can facilitate processing of adpositions and enable to perform better in the generalization task when they are exposed to novel Japanese words/sentences. I wonder whether the authors considered (benefits of variability in) accent adaptation as a confound in their design.*

We believe that the use of the noun training task described above, in combination with our choice of using English cognates plus the fact that they view a demo using the same voice** heard in training before they begin the experiment, should provide participants in both conditions with plenty of opportunity to adapt to the single speaker.

The paper of van Heugten and Johnson (2014) suggested by the reviewer is interesting, however we think it is hard to say how relevant findings with infants (max age: 15 months) are to 7-8 year olds. We also note that in experiment 4 and 5, van Heugten and Johnson (2014) show that, even with infants, prior exposure in the form of a short story read before testing made infants able to overcome difficulties. As mentioned above, we provide prior exposure in the form of the noun practice task and the demo.

**Note that there was an error in our previous manuscript where we said that we recorded two speakers (page 34, line 946 in the old manuscript). This was a hang over from the fact that we did (partially) record two speakers, but in the end we only used one of them across all tasks in the demo, in both noun practice tasks and in learning/testing. Therefore we confirm that we have only one (female) speaker and that we have removed the reference to the male speaker in the text.

4. *I see that the authors provided noun practice with isolated Japanese nouns; how many trials did it contain? Did all participants display accurate word recognition?*

We had missed this info in the previous version of the manuscript, thank you for spotting that. Each noun practice task has 8 trials, one per noun in each of training and testing. In pilot 1, by mistake in session 2 noun practice had 10 items and thus 10 trials instead of 8, we had reported this info in page 41, in the footnote. The majority of the children are at ceiling in these tasks in both of the pilots we have run so far (Pilot 1: mean = 0.91, SD = 0.28; 91% of the children scoring above 80%. Pilot 2 from previous manuscript (conducted online but now not included due to the further changes in methods): mean = 0.9, SD = 0.31, 9 children out of 10 scored above 80%) and regardless of their learning condition (Pilot 1: HV = 0.91 (.25); LV = 0.91 (0.28);) suggesting that children are able to recognize the English cognates – at least in isolation. We provide this info now in the manuscript for pilot 1(pag 66-67, lines 1721-1725, table 7), we will add the same info for pilot 2 when we will get the data (we have proposed a new grid design, see point 5 and 10 in this letter), as reported in pag 73, line 1866.

5. *Related to the first point, and more concerning is the use of unique sentences in the low-variability group. Given that there were only 2 unique sentences per spatial word, I wonder whether children performed the task as expected (i.e., extracted the adpositions from the speech stream and learnt them) or whether they simply memorized the whole sentence and performed the task as a function of the first word they have heard in a sentence. That is, if they hear a word shampoo at the onset of a sentence and it has been associated with the moving object being placed above, they can simply learn an association shampoo onset – object above, without relying on the target adpositions per se*

Unlike in pilot 1, in the new version of the experiment (which you can play in the new demo here) we show children not only the objects heard in the sentence, but also distractors for both the moveable and the “anchor” objects. This means that, at least in the early stages of learning, children need to listen long enough to the sentence to know which pictures are involved, and where they go, regardless of condition.

At the same time, it is true that, after a while, participants in the low variability condition might map the whole sentence, or any part of it (including the individual nouns), to the above/below meanings. However, this is part of the point of the difference between HV and LV. In the model, we see that in the LV condition the above/below meaning does indeed get associated both with the particular nouns, and with their combination, and critically this happens at the expense of learning the mapping with the postpositions. We discuss this in our simulations section.

6. *In addition, when applying this strategy, the low-variability group might not need to recognize the Japanese cognates in running speech per se (which distinguishes them from the high-variability group); this can penalize them in the generalization task, as they would have not learnt to recognize/adapt novel Japanese words in running speech.*

See our points above (point 2) about how we are mitigating against the possibility that differences in identifying the nouns might underpin differences between conditions.

7. *The game provides an option to replay the sentence (multiple times) by clicking on a speech bubble. Related to the first point, is it possible that participants in the high-variability group used the replay option more than children in the other groups, which would lead to an increase in their overall exposure to the*

stimuli? Did the authors examine the number of replays in each group and whether it was related to their performance?

The reviewer raises a valid point about how our training set-up could lead to differences in how often the children heard the sentences in the two conditions.

In fact, in the previous proposed version of the experiment (and pilot 1), there were two ways in which having more difficulty in training might have led to more opportunities to hear sentence stimuli:

1) They might have made more use of the replay button (as the reviewer notes)

2) When they made an error, part of the feedback was that they heard the sentence again, as well as seeing where they should place the object.

While both of these are potentially problematic, on reflection, we are actually more concerned about (2) since it means that children in the HV condition were certainly getting more input. Looking at the pilot 1 data, there is no evidence that children who heard these "extra" stimuli did better – Indeed the evidence is the contrary: Performance in training and testing are correlated for these children, meaning that children who got more input through making more errors in training, actually did worse at test. On the other hand, knowing that children in one condition get substantially more exposure seems like a confound it would be better to avoid. For that reason, we have decided to change this in the methods for the new study – when children make an error, although they will see the correct move, they will not hear the sentence replayed.

Turning to (1), unfortunately, we did not record how often the children used the replay button in the pilot experiments. Anecdotally, the last author was present when many of the pilot 1 children were tested and does not recall that they used it very often beyond the first few trials. The reason we included this button was that Hsu & Bishop (2014) had this replay option, and we thought that if children were distracted, they might need another chance to hear the stimuli, particularly in the early stages. This may additionally be more necessary given the changes which we have proposed to the new version in terms of having more foil nouns compared with pilot 1 (see point 10 below). We are thus not sure that it is wise to drop this option. Instead, for the new version of the methods we will record the usage of the button so that we have some indication of how often it is used by each child, allowing us to check how this relates to performance and, critically, whether it differs between conditions.

8. *I find it a bit problematic, given the focus of the manuscript, to exclude participants who display consistent floor (average performance at chance level, i.e., 50%) or ceiling effects (average performance above 90%) during training in both sessions. If one of the conditions is intrinsically more difficult than the other, then applying this criterion might artificially increase the mean for a difficult task and decrease the mean for an easy task. I understand the authors' concern that parents might interfere, but this can also happen when children perform "as expected". When we look at Pilot 1 data, run by an experimenter, we can see that many children remained below the chance level even after the second session, in particular in the high-variability group. Did the pilot data revealed statistically significant differences in the number of participants displaying floor and ceiling effects between the conditions? How much data was concerned?*

After careful consideration of yours' and other reviewers' comments we have decided to remove this exclusion criterion and have it based only on the noun practice tasks: We will reject children whose performance is below 80% across sessions. Note that 91% of the children in pilot 1*** and 90% of the children in pilot 2 (soon to be replaced by a new pilot 2) met this criteria.

*** We have also repeated the analysis for pilot 1 applying this exclusion criterion, and results do not change in any meaningful way (beta for learning condition at test = .54 vs .59, beta for learning condition at training = 1.37 vs 1.34).

9. *Likewise, I find it problematic that trials in training whereby children pick up either the wrong moveable object, or they position it in one of the distractors cell will be removed. Doesn't this reveal that children were not able to accurately segment the sentence or/and recognize the Japanese words or/and learn the postposition? These errors inform us about how difficult a given condition is and its learnability (8% of trials, as revealed by Pilots, seems considerable to me);*

Our key interest is whether participants have learned the meanings of the constructions involving the postpositions and how this differs across conditions. For that reason, we prefer for our main analyses to focus on trials where the participant has identified the correct nouns and ask, given that, do they know the correct meaning. In the way we propose to look at this in the new manuscript (see further next point), this means we will only analyse responses from trials where the participant moves one of the two objects mentioned in the sentence, and places it either above or below the other object mentioned in the sentence. Another benefit of this is that it makes our main analyses identical to those in pilot 1, where there were no distractor nouns, making our power simulations more relevant.

On the other hand, we fully agree that there could be useful information in the "excluded" moves. For this reason, we now plan to also look at whether the proportion of "excluded" trials differs between the conditions (sanity check section, pag 49-50, lines 1310-1342). If it does, this may speak to the reviewers prediction that children in one condition might find these stimuli more generally difficult, whilst allowing our main analyses to focus on the learning of the construction.

10. *in addition, it is not clear to me how removing these erroneous trials makes the task comparable to a two-alternative forced-choice task. Maybe the authors could comment on this?*

We acknowledge that our previous manuscript was confusing on this point. This was in part because chance was computed differently in pilot 1 than in the main experiment, and in the main experiment training and testing were different. We have now simplified things by making the grid set-up for training and testing identical in the new design, and matching the chance level between pilot 1 and proposed experiment.

We have now added a more detailed explanation of the type of moves which it is possible for participants to make using a schematic representation (figure below, and page 38-39, lines 1009-1029, figure 8). This makes clear that – for training and testing – there are four possible moves which will be included in our main analyses, once we have removed all trials where they had not identified the two nouns correctly.

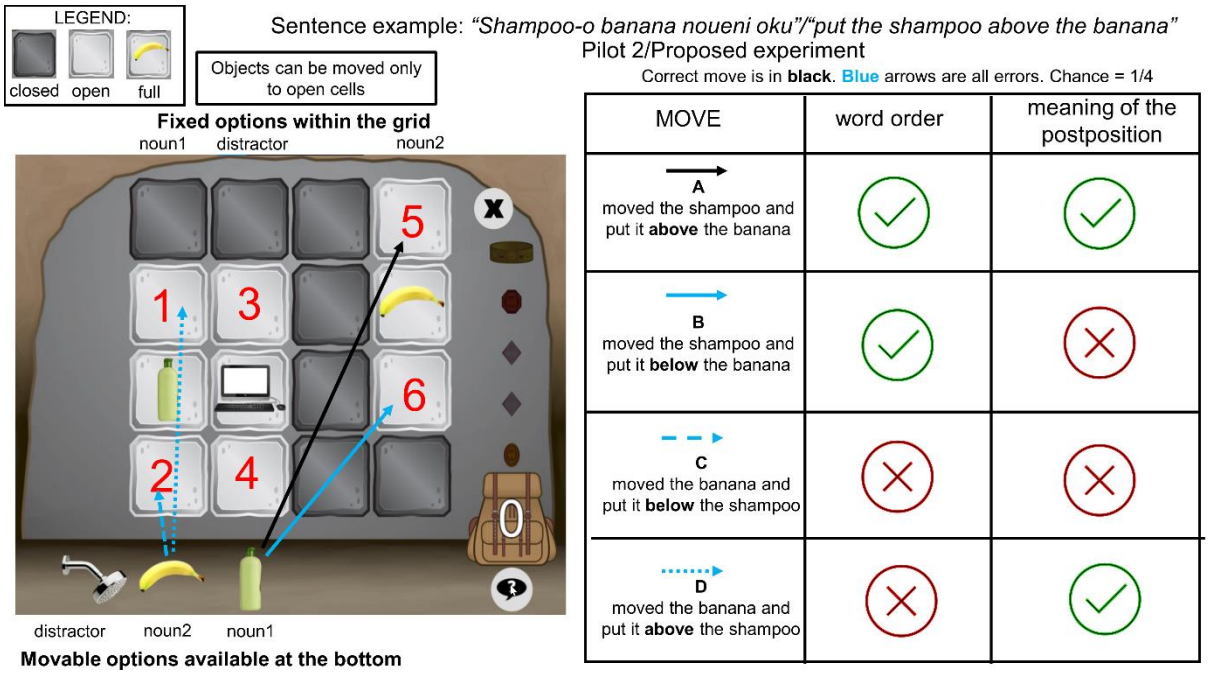


Figure 1 – An example of the stimuli configuration for a sentence like ''Shampoo-o banana noueni oku''/''put the shampoo above the banana''. The game allows the player to place a picture only in open cells numbered from 1 to 6 in red. In this trial, the shampoo is the movable object (noun 1) while the banana is the anchor object (noun2). The correct move is represented by the black arrow. Blue arrows are all errors. The arrows indicate the four moves which children could make which would indicate that they had identified the nouns within the sentence – I.e., that they realize that one of the two nouns in the sentence has to be placed above or below the other noun made in the sentence. All other possible moves (e.g., moving the banana to square 6, i.e., below the computer) will be excluded from the main analysis, making the likelihood of making the correct move ¼. On the right side of the picture, all moves have been classified based on whether they speak about children's ability to learn word order, the meaning of the postposition, or both.

11. *The authors do not include nouns as a random factor in the model. Is it possible that some items, e.g., "ice-cream" perform better than others?*

Thank you for raising this point. With our pilot data, we have tried adding items as a random factor. Unfortunately, this raised convergence and singularity problems. Given that items are randomised across participants, we feel that it is not necessary to keep this variable within the model. However, we think that other readers may also question this choice and so we have added a footnote on page 64.

12. *For the third hypothesis, is it possible that the word for above in Japanese is acoustically more salient (with the three vowels following each other) as compared to the word for below?*

This is an interesting possibility – thank you for the suggestion. It could be that this underpins higher performance with noueni, as well as/instead of the fact that the ordering is more iconic in those sentences. We now discuss suggest both of these possibilities in the manuscript where we present that hypothesis, and acknowledge that our design can't distinguish between them (page 30, lines 823-830). We don't consider this problematic since the primary goal is to look at the effect of learning-condition, whereby postpositions are considered together. Nevertheless, we predict the

differences between postpositions on the basis of our pilot, and we plan on reporting and discussing the result further in our general discussion.

13. *Do hypotheses 8 to 10 refer to the training data or to novel noun trials?*

Hypotheses from 8 to 9 refer all to the training data; however hypothesis 10 relates to a possible relationship between performance in training and performance at testing. Each of these hypotheses is organized within a subsection: 8 and 9 within “predictions for training data” (pag 30) and 10 in “predictions for relationship between training and test” (pag 32). We acknowledge though that this might not have been easy to read, so we have now made the subtitles bigger and in bold.

14. *Minor comments:*

- *L 17: an abstract => remove “an”*
- *L 267: a=> at*
- *L 285: countering => encountering*
- *L 497: that predicts=> predict*
- *L 603: If my understanding is accurate, the “set of 10 unique sentences with frequency of 8 and a set of four sentences with frequency 1” result in 84 sentences for the skewed condition, whereas there are 56 sentences in other conditions but also on figure 2 for the skewed distribution. Please check for consistency.*
- *L 757: remove one “without”*
- *L 1113: in the in novel nouns test => remove second in*
- *L 1235: to corresponds => to correspond*
- *L 1236: referred to*
- *References:*

  van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. Journal of Experimental Psychology: General, 143(1), 340–350. https://doi.org/10.1037/a0032192

Thank you – these have now been corrected.

Reviewed by Caroline Rowland, 22 Jan 2022 16:14

15. ***Overview.*** *This is an very comprehensive stage 1 proposal that combines computational modelling and pilot data to motivate a study of the relationship between input structure and generalisation in an online experimental paradigm with 7-8 year old English speaking children. The authors will test the hypothesis (generated from their theory of discriminative, error-based learning) that children will learn the meaning and use of Japanese spatial adpositions more effectively when there is more variability in the use of the nouns within spatial sentences. They also propose to test a number of secondary hypotheses; most notably, an empirically generated hypothesis that skewed distributions might be as good (or even better) for learning generalisations than highly variable input. The report satisfies all the necessary criteria in my opinion; the authors have done an excellent job. Below I summarise my comments under the 5 headings/areas suggested in the Guidelines for Reviewers, before finishing with some more general comments.*

    1A. The scientific validity of the research question(s).

    The research question is scientifically valid and is detailed, with sufficient precision as to be answerable through quantitative research. The authors motivate their theoretical perspective with a literature review, and with a computational model. The study proposed falls within established ethical norms for working with children of this age.

    1B. The logic, rationale, and plausibility of the proposed hypotheses, as applicable.

    The proposed hypotheses are coherent and credible and are very robustly motivated. The authors motivate their primary hypotheses with a detailed, evaluative literature review, and a computational model. Secondary hypotheses are motivated empirically (on the basis of previous studies and/or pilot data). Both types of hypothesis are stated precisely and are sufficiently conceivable to be worthy of investigation. They follow either directly from the research question, or indirectly via empirical evidence (in the latter case, the analyses will yield important additional information that might lead to modifications of the theory). The pilot studies (pilot 1 and 2) are also well designed and well explained (though I have one point regarding the chance level of 25%, which I address under 1C below).

Thank you, Prof. Rowland. We are glad that you found our manuscript interesting and our work valid.

16. *However, I would like the authors to address one point here regarding the statement (page 31) that they also plan to include measures of individual differences (e.g. attention, vocabulary) for exploratory analysis. I recognise that these are exploratory analyses, but the authors should motivate them in some way – what factors will be assess here, what relevant information might the tasks yield, why are individual differences of interest here etc. In addition, these task are not mentioned at all in the methods section (see point 1.c below).*

Thank you for spotting this. We have now decided to drop most of these tasks because we do not have clear predictions.

The only task that we have decided to include is a working memory task (page 41, lines 1063-1070) which Dorothy Bishop and Adam Parker use in their replication of the Hsu & Bishop (2014) study, which we discuss in the paper, and it's available on OSF. They are using the same task in an online setting with a group of adolescents. Thus, although this task is not validated for this age group, we think it could be useful to have the data for future comparisons with their study. In addition, given our discussion of how skewed input might be particularly helpful due to capacity constraints, we think it

could be interesting to see if there are interactions between this working memory measure and the difference conditions, particular the skew and HV conditions. We have added a comment to this effect in the manuscript (pag 33, lines 906-911), however we keep it brief since we do not have clear priors for our predictions and the analyses are going to be exploratory.

1C. The soundness and feasibility of the methodology and analysis pipeline (including statistical power analysis or alternative sampling plans where applicable).

The study procedures and analyses are incredibly well described and are valid. Critical design features such as randomisation, rules for exclusion etc, are present and fully explained. Please note that I have not conducted Bayes analyses myself, so my knowledge of what needs to be considered is purely theoretical. Bearing in mind that caveat, the proposed sampling plan is rigorously described, and the thresholds for evidence at different levels (strong, moderate etc) for H1 and H0 are clearly described.

However, I have three points for the authors to consider:

17. *As mentioned above, the authors state on page 31 that they also plan to include measures of individual differences (e.g. attention, vocabulary) for exploratory analysis. There is no mention of these tasks at all in the methods section. If these tasks are to be included, please add the usual methodological details (e.g what the tasks will be and how the data will be collected).*

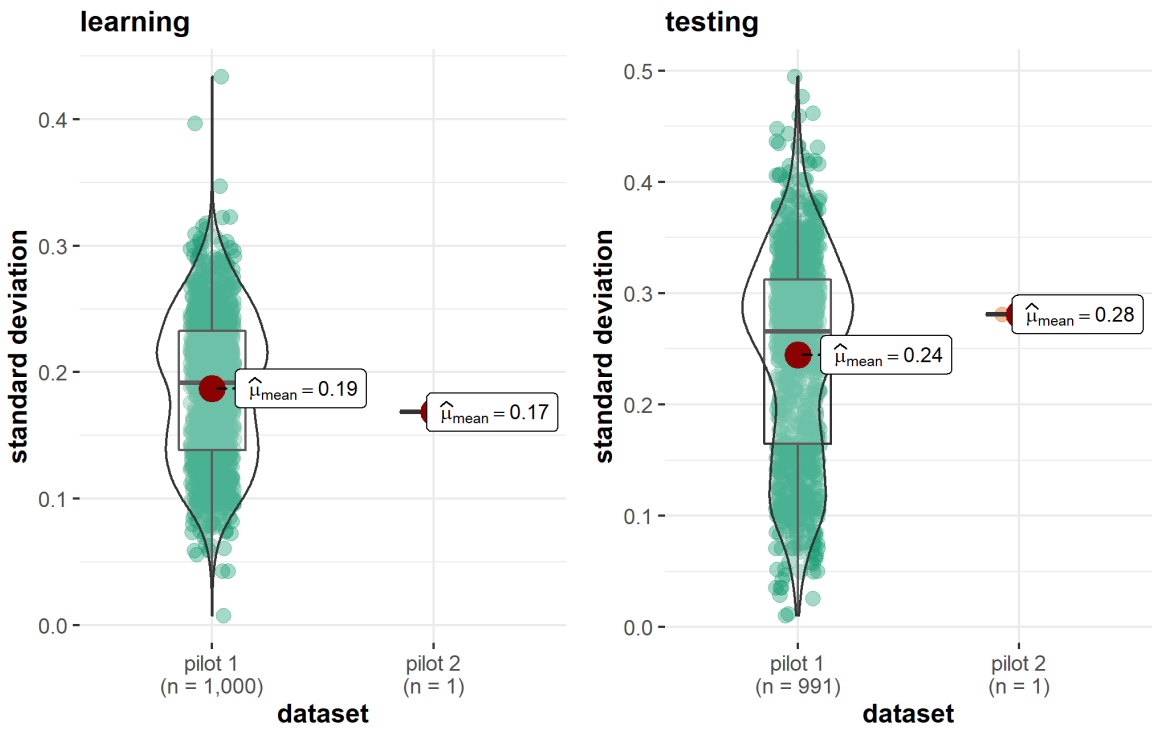
Please see our answer to point 16 above

18. *Motivation for the choice of statistical priors. Their priors are defined on the basis of effect sizes taken from the pilot study, which was conducted in person in the children's schools. I think there is now strong evidence that data collected online tends to be nosier – and effect sizes smaller – than data collected in person. This is particular the case with studies with children, and even more so when in person data was collected in a structured environment such as a school, where there are minimal distractions. If their online study yields substantially smaller effect sizes than their pilot data, will their study still be adequately powered to find strong/moderate evidence for H1 and/or H0 for all their hypotheses?*

We acknowledge that this is a possibility, although concerns about online experiments being noisier than in lab are mainly related to studies that measure RTs (e.g., Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W., 2020, *PeerJ*.; Anwyl-Irvine, A., Dalmaijer, E., Hodges, N., & Evershed, J., 2020, *BRM*). In our experiment, we measure performance as proportion of correct/incorrect responses. Studies comparing in lab vs online experiments whose dependent variable was accuracy/performance demonstrate that while there might be noise due to replication of the same experiment twice, results are consistent between settings (e.g., Horton, J. J., Rand, D. G., & Zeckhauser, R. J., 2011, *Experimental Economics*; Dandurand, F., Shultz, T. R., & Onishi, K. H., 2008, *BRM*).

In our study, we can compare the amount of variance for children in the online pilot that we have done so far (now removed from the current manuscript due to the changes in methods) and children in the pilot 1 which was conducted in school (focusing on HV condition, so that it is matched to the online pilot). The figure below shows the standard deviation for the N=10 children in pilot 2 in training and test versus the distribution of standard deviations for around 1000 unique random samples of 10

children drawn from the HV condition in pilot 1 for each of training and test. It can be seen that the standard deviations are in the same ball park.



Regarding the choice of priors, we follow advice that the best approach is to use a prior that stems directly from actual data (like we have done here), without tweaking it a posteriori because of the risk of a possible bias. As reported elsewhere (e.g.,: Evans, M., & Guo, Y., 2021, *Entropy*), choosing a prior to minimize one type of bias simply increases the other. We note that if our prior is indeed optimistic, as Prof. Rowland suggests, then this biases us towards finding evidence for the null over H1 (that is, for Bayes factors, the conservative choice is to set a higher prior).

19. *Chance level. On page 35, the authors state that they will remove trials in which children make 'illegal'' moves (e.g. placing the object on a distractor cell), so that chance level is 50%. However, I'm not sure this is right. Even on trials in which children make legal moves, they still have the possibility of making an illegal move (placing an object on a distractor cell, or of choosing a distractor object). So even on legal moves (which are included in the analysis), chance is less than 50%. This doesn't (I don't think) have any profound implications for the analysis because none are comparing performance with chance (though authors should check this). But either way, if I'm right, the authors need to:*

    - *calculate the actual chance levels across trials and state this in the paper.*
    - *Or remove distractor objects and cells*
    - *Or (and this is my preferred option) make it impossible to place objects on distractor cells (i.e. make distractor objects non-movable and make objects ping back to their original position if you try to place them on a distractor cell).*

We have realised with yours and the anonymous reviewer's comment that our explanation of how moves would be removed from the analyses, and thus chance levels was not clear enough. We now provide a more detailed explanation of this (page 38-39, lines 1009-1029, figure 8) and within this letter at point 10. Note that these refer to the new version of the game (available at this link – depending on the internet connection the game can take up to 2 min to load during which you will see a blank screen) that we now use in response to the suggestion of the anonymous reviewer about including distractors that allow us to test whether children have learned the word order, as well as the meaning of the postpositions (point 9 in this letter). In this new version, for the main analyses, "chance" in both training and testing is 25%. The reason for this is that prior to doing our main analysis, we will remove all of the moves that indicate that they haven't identified the two nouns involved in the sentence (so remove all those where they don't move one of the two objects in the sentence either above or below the other). Thus for this main analysis, we are looking at a dataset where only four moves occurred, making chance ¼.

As explained in the manuscript (pag 73-74, lines 1882-1895), the key reason that we have introduced the distractor nouns is to mitigate against the possibility that children in the LV condition do not have to listen to the sentence and instead base their response on the scene, which might happen if there is no variation in the scenes, which was the case in pilot 1. On the other hand, we want our main analyses to reflect the core learning of the construction, rather than ability to identify the nouns, hence the fact that we remove these from the main analyses.

In terms of Prof Rowland's specific suggestions, note that distractor objects placed on the grid are already non-movable. However we don't want to make the movable objects ping back to their original position for two reasons:

- The first one is practical: Currently, moves that results in objects bouncing back at their starting position are those related to illegal moves. For instance, if children move around objects outside of the grid, or they try to place it in a closed cell within the grid. These moves do not follow the game's rules. Extending this type of behaviour to include moves which involve the distractor column in training could be confusing.
- The second one is theoretical: As just noted, our goal was to encourage children in both conditions to pay attention to all of objects in the scene and if they quickly learn that some objects are not movable they may be less likely to do this.

20. *Note that this issue also applies to the analysis for pilot 1, where chance is set to 25% for the same reason (page 58). Again, I don't think this is right; i.e. removing trials from the analysis where children make illegal moves doesn't make any difference to the chance level on legal trials. But please tell me if I'm wrong about this – I may have misunderstood what the authors did here.*

We understand the confusion – we could have been clearer. We have made now a schematic representation of legal and illegal moves in pilot 1 that we have added in the manuscript as well, pag 66, figure 13. Note that we have also changed the names of the response types by using letters instead of numbers. As one can see, by not considering illegal moves there are only 4 possible moves (children can move only one object per attempt), and only one is correct (highlighted in black).

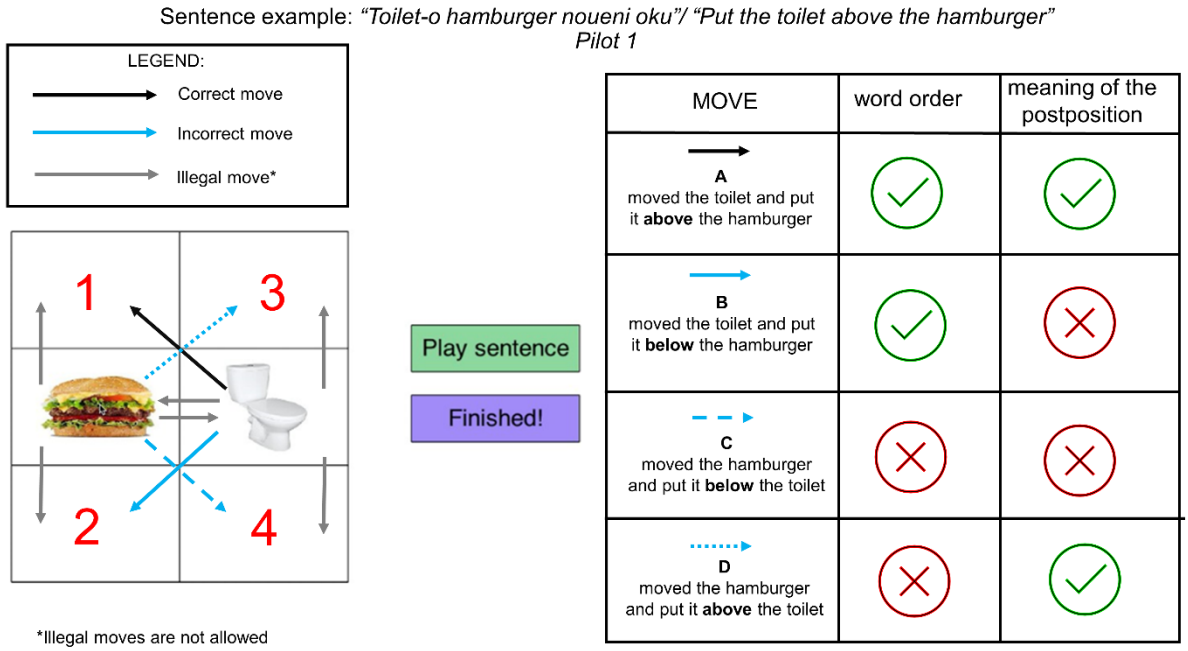


Figure 2 Illustration of the possible (i.e., legal) and impossible (i.e., illegal) moves in pilot 1 for the sentence: ``Toilet-o hamburger noueni oku''/``Put the toilet above the hamburger''. In this trial, toilet is the movable object and hamburger is the anchor object. All blue arrows are legal errors, the black arrow is the correct move. We removed illegal moves (arrows in grey, e.g., toilet moved to square 4) from the first batch of data collection and programmed the experiment not to allow them anymore for the remaining participants. Among legal moves, the blue arrow from the toilet (i.e., the movable object) is a type \added{B} error and those from the key constitute type \added{C} and \added{D} errors. The schema on the right side of the picture classifies all legal moves based on whether they are informative about word order, meaning of the spatial postposition or both.

> ***1D. Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.*** *The protocol certainly contains sufficient detail to be reproducible and ensure protection against research bias, and specifies precise and exhaustive links between the research question(s), hypotheses methods and results. The design summary table is very useful. Please note that some reviewers might state that they find the introduction section overly long. It is, indeed, very comprehensive. However, I appreciate this. It lays out, very clearly, the authors; theoretical perspective, the learning mechanism they are proposing, and neatly evaluates all the relevant previous literature. As many of us are now arguing, there aren't nearly enough papers in the child development literature that really get to grips with potential mechanisms of development; i.e. we have too few papers that accurately explain, in detail, how learning processes might work. Thus, I find it admirable that the authors have prepared such a careful, detailed review. There is only one sub-section I might shorten, which is the one describing the study by Hsu & Bishop (2014). However, even here the detail can be justified given how closely that study relates to this one.*

We agree with Prof. Rowland's point of view, and we thank her for providing her personal perspective on this issue.

21. ***1E. Whether the authors have considered sufficient outcome-neutral conditions (e.g. absence of floor or ceiling effects; positive controls; other quality checks) for ensuring that the obtained results are able to test the stated hypotheses or answer the stated research question(s).***

    The proposal contains all the necessary data quality checks (though please note my worry about the statistical priors detailed in 1B above is also relevant here). Proposed statistic tests are appropriate and outcome neutral. The pilot data suggests that there are unlikely to be floor or ceiling effects. Positive controls are appropriate (high, low, skewed variability input).

We have addressed the concerns about data quality checks in point 16.

22. *Page 10, footnote: "From a theoretical perspective, we do not believe there is any good reason to expect transfer to new constructions..." I don't quite agree with this – there is good evidence for construction-general transfer in some circumstances (e.g. Abbot-Smith and Behrens' wonderful construction conspiracies'' paper (2006)).*

Thank you for pointing us to this very interesting paper. Yes, in light of this evidence and on further reflection this may have been too strongly stated. We have reworded this as follows (page 10, footnote):

"From a theoretical perspective, although transfer across related constructions does happen in some cases (Abbot-Smith & Behrens 2001) we do not believe there is any good reason to expect strong transfer to new adpositions in this paradigm"

23. *Page 12: Both paragraphs on skewed distributions. I found it really hard to follow these two paragraphs; Paragraph 1 seems to be saying there is a skewed distribution in natural language, and paragraph 2 contradicts that. I think I know what the authors are saying but it's a bit confusing. Can they rephrase?*

After careful consideration, we have decided to rephrase the paragraphs in order to make them clearer. We have highlighted all the changes and novel additions in red.

24. *It would also be useful to give a short definition of a 'geometric distribution' in the text, so readers don't have to read the footnote to understand what it is (footnotes should probably just include additional information, not information essential to understanding the main text).*

In the novel version of the paragraphs we now provide a shorted definition of the geometric distribution without using the footnote (which we have now removed). We thank Prof. Rowland for the suggestion.

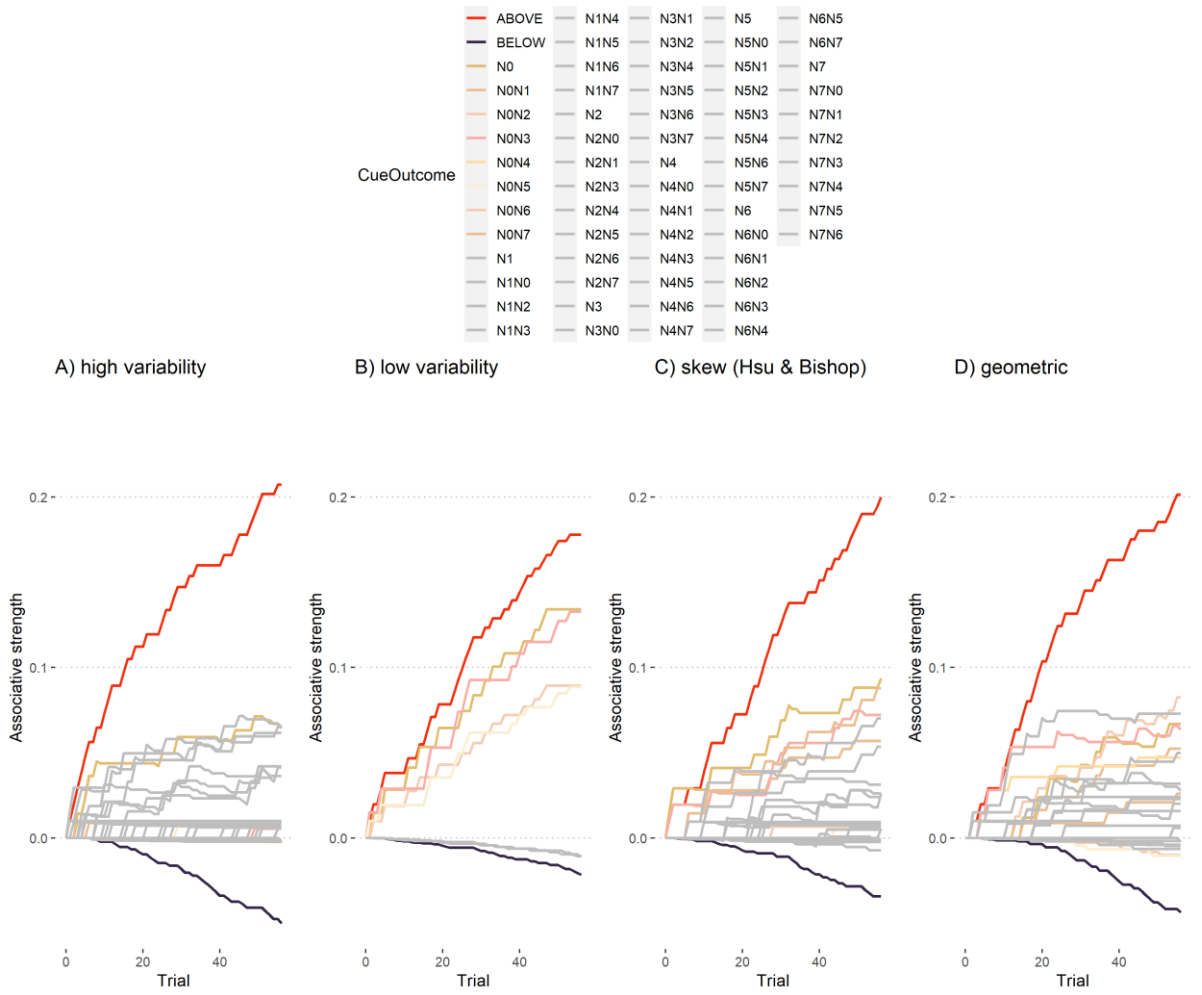
25. *Page 18: "word order is not captured by the model'. I wondered what consequences this had for learning, and for the comparison with real children, since children certainly do have access to word order cues. If word order *was* captured by the model, how would this change the pattern of results (if at all)? Could the authors speculate here?*

We agree that this point needs some discussion. Children do get word order cues and likely have biases based on linear order, for example they likely attend more to the first word in the sentence.

This last point can be addressed by increasing saliency for the first noun in the sentence, so that noun 1 is more strongly weighted than the other cues present in the sentence (both nouns in second position and the postpositions per se). We have tried this and the results are in the figure below. We have increased saliency for the first noun by 50% of its default value (i.e., 0.15 instead of 0.1). Of course, the magnitude of the increase is arbitrary without having an a priori hypothesis, but for illustrative purposes we can hypothesise that its saliency is 50% higher than the saliency of the other cues (postpositions and nouns in second position).

As one can see in figure below, the increased saliency for the first noun causes a spike in the associative strength compared to the nouns in second position (this is easier to see in the LV condition where we have only 2 types of sentences associated to the outcome and we can see how the learning trajectories of the nouns bifurcates into two distinctive paths based on whether they are in first or second position), but it makes no noticeable difference to the learning trajectories of the noueni/noshitani postpositions and thus makes no difference in our key comparison of the differences in these trajectories between conditions.

This suggests that for this particular type of learning, the saliency of the first noun caused by hearing the nouns in a particular ordering is not key for the model. Of course there are other features of word order which might be important. We therefore find it interesting that, even without including word order, the model captures the difference between conditions seen in our pilot 1 data. Our new proposed design will allow us to further establish more precisely children's learning of word order as well as the meaning of the postposition (see "Follow-up analyses" at pag 54-56), and we will add these consideration in the Discussion, at stage 2, when we can also compare how the model's predictions fared in comparison to the children's performance.



26. *Page 27: difference between simulation results (no benefit of skew) and empirical results (benefit of skew). I did wonder what the simulation results looked like earlier in the learning cycle. One possibility is that the simulations have just learned the generalisation much better than children by the end of the learning cycle. So if we want to replicate empirical results (especially from children with DLD) we might want to look at the data earlier in the learning cycle of the simulation. If we administer the test session earlier in the model's learning cycle, is there any evidence of a skew advantage?*

We did consider whether there would be differences earlier in learning, but this is not the case. This is evident from figure 3 which shows change over time: the learning trajectories show no strong evidence of differences between the HV and skew conditions for the key mappings between the postpositions and above/below. We will further discuss how the modelling fares against the human data once data collection is complete in the Discussion, at stage 2.

27. *Page 60; Table 7. Please add a description of the three "response types" to the label of table 7. I was initially confused until I realised these referred back to the "four possible moves" described on page 58*

Prof. Rowland is correct, they do refer to the moves on page 58 (now page 65 and 66, and figure 13). We now clarify this in the caption of the table – thank you for the suggestion. Note that we refer now to those moves with the letter of the alphabet (A, B, C, D). We have highlighted this change in the text and uniformed the moves' nomenclature across the proposed methods and pilot 1.

28. *Throughout, but especially in the sections describing the pilot, two different labels are used for the skewed input: 1) skew-bimodal/skew and 2) exponential/geometric. This can make the paper a bit different to parse (especially on pages 24 and 25 where the labels on figure 4 refer to skew-bimodal and exponential conditions, but the description of the figure in the text uses skew and geometric labels).*

After consideration we decided that it is better to use the more familiar term "exponential" rather than "geometric" in this paper. We therefore now only mention the term "geometric" once, when we point out that strictly speaking exponential distributions in linguistics are geometric since we are measuring something discrete rather than continuous (the same incidentally is true for covid growth, though this is always described as exponential – hence why we think this more common term is probably better to use here). We have also included a couple of sentences which explain the relationship between power law distribution and exponential distributions.

29. *There are a number of typos throughout so a good proof read would be useful. I have not listed them here because of the time that would take. If the authors would like to see these, please can they send me an editable version of the manuscript (e.g. googledoc or overleaf) and I can use track changes to point them out. (NB noueni and noshitani are sometimes italicised and sometimes not).*

We very much appreciate Prof. Rowlands kind offer to identify these in overleaf and we are happy to give her access to our Overleaf via her email of preference for this updated version. We have done our best to correct all the typos and mistakes which we were able to see. In places, we preferred to use the italicised forms to give emphasis. We hope Prof. Rowland agrees that this is more a matter of style rather than substance.

Reviewed by Julien Mayor, 21 Jan 2022 15:21

30. *In their Registered Report, Viviani, Ramscar & Wonnacott introduce a study in which they will examine the relationship between input structure and generalization in a training study in which 7-year olds learn spatial adpositions ("above" and "below") in an unfamiliar language (Japanese). In that study, children take part into a computer game and are instructed to move objects above or below other (fixed) objects, using instructions in the unfamiliar language (Japanese). Children are either exposed to a low variability condition (in which nouns within spatial sentences are frequently repeated), a high variability condition (in which there is more variability in the use of nouns) or a skewed distribution condition (in which a subset of items are repeated while others are presented fewer times). Using computational simulations, using a Rescorla-Wagner model, the authors predict superior performance in the low variability condition during learning, while children in the high variability are expected to perform better than those in the low variability condition at test (with novel nouns). The authors will examine how children exposed to a skewed distribution will perform during training and at test, in comparison to the other two conditions.*

    First of all, this is a privilege to be offered with the possibility to provide input to a study before it has been conducted. And my first comment is to laud the authors for their impressive study and manuscript. The "game" in which the study is embedded is stunning, the statistical pipeline flawless, the piloting phase is extensive and the reading flows extremely nicely. This is one of the most thorough Registered Reports that I have ever had the honour of reading, and it is a humbling experience to be asked to comment on it. I will, however, try my best! Let me start with what I think is the most important point first.

Thank you for the kind comments and the time dedicated reviewing our manuscript.

31. ***Word Order.*** *The training regime does not assess word order, as the first noun in the sentence is always going to be associated with the object to be moved "above" or "below" the second item. In doing so, children are never offered the possibility of distinguishing between "X above Y" vs "Y above X". This explains how children (may) perform better with one adposition ("above") than the other ("below"). In fact, children can be 100% correct with one adposition without ever listening to the sentences (always putting the moveable object on top of the fixed object). So, wouldn't it make sense to adopt a procedure in which both X and Y can move respectively to each other (as it is the case at test)? This way, sensibility to word order could be assessed – which seems essential if one would want to evaluate whether children understand the meaning of adpositions ("X above Y" is not equivalent to "Y above X").*

The reviewer is correct that our proposed method didn't test word order, unlike pilot 1 which it did. We did this because when we probed the pilot 1 results, we found that children in the two conditions were equally good at knowing that the first-heard noun was the one that should be moved (this info can be found in the novel version of the manuscript at pag 71, table 9, which shows that the majority of the errors are of the B-types with no differences between HV and LV), and the difference was driven by whether they knew to move it above/below the second mentioned noun (see also pag 69, lines 1769-1774).

We had therefore thought that it was reasonable to use a procedure which did not assess word order in this new versions, since it made the task a little easier and it might offset some other changes that decrease the difficulty. It also made children's test closer to what the model is doing (see response to next point below). However, on much (!) further reflection, we have decided that the reviewer, is right: knowledge of the word order is key to learning the construction, and so it is important that it is reflected in our empirical test.

We have therefore changed the paradigm such that if the two nouns are X and Y, the child will be able to move either of these with respect to each other. This is done by having both X and Y appear both in the set of "movable" objects and within the grid where they serve as "anchor" objects as can be seen in the figure 6 on page 37 of the manuscript (To try the new design, please click here – note that it can take up to 2 min to load the experiment during which you may see a blank screen, simply wait a few minutes). We provide further explanations in the grid configuration subsection, pag 36, lines 995-1001. Given that this change is substantial, we plan to run a further pilot to check that children can still meet baseline's measures of performance in the new paradigm (when complete, this will replace pilot 2 in the previous manuscript).

32. *Relatedly, the computational model does not capture word order effects either. Weights associated with each item in the sentence are summed, independently from the order in which they appear. In doing so, the model does not either distinguish between "X above Y" and "Y above X". As such, one may ask whether the model appropriately captures the acquisition of the meaning of adpositions. Of course, one may argue that, precisely, neither children nor models need to process word order so, from that perspective, the model may accurately capture children's behaviour. But, in sum, neither children nor models are really assessed on their comprehension of the meaning of the adposition, as they are never assessed on their capacity to distinguish between "X above Y" and "Y above X".*

There is currently not clear consensus on how to add word order into this type of model, and we feel that attempting this is beyond the scope of the current paper. However, if we try to consider word order by, for instance, increasing saliency of the first noun we find only negligible differences and limited to noun learning, rather than to postpositions. We have tried this approach in point 25, in relation to Prof Rowland's comment.

Instead, in the paper, we acknowledge that the model as presented is a simplification which doesn't incorporate this aspect of learning the construction meaning (pag 18, lines 594-596). In the end, all models involve simplifications and we think it is interesting that – even given this particular simplification – the simulations nevertheless captured the difference that children showed in the HV and LV conditions in the pilot study. We of course hope that this will still be the case with the new design. The proposed changes now mean that children do have to learn the word order, and so in that sense differ from the model. However, in addition to the main analyses, we propose a follow up set of analyses (pag 55, lines 1488-1505) where we will only look at trials where response indicated knowledge of the word order – i.e., trials where they moved the picture corresponding to the first noun and placed it either above or below the picture corresponding to the second noun (so trials of type A and B in figure 8, pag 39). This isolates testing of the meaning of the postpositions and thus is closer to the model.

33. *Another issue is that, at test, the authors suggest a baseline performance of 25% - when children can move, relatively to each other, both objects X and Y. I wonder if children have any way of distinguishing between two interpretations of the instruction sentences (in Japanese)? They can either (a) understand that they need to, e.g., move object X above object Y (in which a strict interpretation of the baseline would be 25% - children get to pick the correct object out of two, then move it in the correct place), or (b) that they need X to be above Y (in which case you can either move X and put it above Y, or move Y and put it below X, resulting in an actual baseline of 50% - as both configurations would be correct with the interpretation of the adposition itself).*

This was a concern in the previous version (pilot 1). However, in the proposed paradigm, in both training and at test, there is a distinct set of movable and non-movable pictures (below and within the grid respectively). Only movable objects can be objects of “put” and so to make a correct response children must put X above Y; putting Y below X is an incorrect response to the command. We acknowledge that in the previous version of the manuscript this info was only implied, in the current version we have emphasized it at page 73, lines 1878-1888.

34. *Relatedly, the difference between the design of pilots (in which both objects X and Y could be moved relative to each other during training, as for testing) and the planned study (in which only X can be moved during training while at test both X and Y can be moved, resulting in fairly different training and testing regime) may not be inconsequential.*

We agree that our previous proposal was rather different to pilot 1. The proposed version is more similar – training and test are now the same in the new version (as they were in pilot 1) and both pilot 1 and the new version allow children to make the same four moves that we have labelled as A, B, C, D in figures 8 (pag 39) and figure 13 (pag 66). The difference between the two is now in what constitutes the set of additional error-moves that children could make which will be excluded from the main analyses.

The new pilot which we propose to run if you, the editor and the other reviewers agree on the changes, will help to establish that the amount of excluded data with the new version is reasonably similar compared to pilot1. We laid out our analysis plan for pilot 2 in pag 72-73, lines 1831-1870.

35. *Other points: In their study, the authors carefully evaluate the impact of input distribution on learning and generalization. The study, however, presents children with just two adpositions. A discussion of the impact of having to learn multiple adpositions in parallel (as it is likely the case in real life) would be nice. Would the relationship between input structure and learning/generalization stand if you were to scale up the number of novel items that children would need to learn? I wonder if this discussion could be supported by computer simulations?*

Thank you for raising this point. It is quite common in language learning experiments of this type that just one or two novel constructions are taught, including those referenced in the introduction (e.g. one construction is taught – In Wonnacott, Boyd Thomson & Golderberg (2012) – and in Cassiender & Goldberg, 2005; in Casasola, 2005; in Goldberg, Casenhider, Sethuraman, 2004; and in Maguire, Hirsch-Pasek, Golinkoff, Brandone, 2008). In each of these studies, we can ask the question of how this would scale up to more naturalistic contexts, and this is of course critical for the implications of the work.

While we agree this point is important, on reflection, we feel that it is not useful to introduce discussion of this into what is already a somewhat lengthy introduction section. However, if we proceed to stage 2, we intend to address this in the General Discussion.

36. *Could it be that Goldilock effects on cognitive load may take place – such that highly variable input just/also provides more engaging stimuli, such that children maintain concentration until test trials (hence better results), whereas low variable input would lead to more rapid "boredom" such that by the end of the training session children aren't paying attention anymore?*

This is a valid concern, although another possibility is that opposite might be true: that in the LV condition because children get always positive feedback they might be more motivated and focused, while in the HV condition since children get more corrective feedback they might lose focus very quickly.

Both situations seemed to us possible and that's why we have extensively piloted our experiment. In pilot 1, we found that 56 trials spanning over 2 sessions did not burden/bore too much children as attested by increasing performance from session 1 to session 2 in both HV and LV conditions. In addition, the idea that children are more motivated in one condition than the other would suggest that we might see stronger performance in both training and testing in that condition, rather than as attested pattern in pilot 1, where LV outperform in training while we see the reverse at test.

37. *Exclusion criteria (p. 48). I'd specify more finely conditions for excluding participants that perform "at chance". The current wording, to exclude children than display consistent floor effects (50%) is not specified enough. What does "consistent" mean? Is 51% acceptable?*

This issue about our exclusion criteria based on performance at training has been consistently found to be inadequate by all reviewers. We recognize this and we have removed it completely. We now exclude children who seem not to be able to recognize the nouns in the noun training tasks as defined by an average performance of less than 80% across the two tasks. See page 48, lines 1295-1300.

38. *Reading the OSF on pilots, it looks like there were no a priori predictions for interactions. I'd highlight in the Hypotheses section when predictions stemmed from the model and when they stem from empirical data using the Pilots.*

We are a little confused by this as within the manuscript in the Prediction section from pag 29 to 33 we motivate every hypothesis with a "Justification" paragraph which specifies when these predictions stem from data, from the model, from the theory and from literature. Note that we separate those when it comes to training and testing data.

39. *The time interval between both sessions will somehow vary. Are you going to control for this time interval? If so, how?*

For practical reasons in terms of recruitment, we do not think it is feasible to fix this gap. However we will have a minimum gap of 24 hours between session 1 and session 2, and a maximum gap of 7 days. We will record the gap and ensure that it is approximately matched. The data will be collected so that we will (or others) can also explore potential impact in subsequent (exploratory) analyses.

40. *A little bit more justification of the age range would be nice. Why 7-year olds?*

There are two reasons for this. First 7/8 is also within the age range of typically developing children that participated in Hsu and Bishop's experiment (6-11), making it easier to make comparisons across the studies. In addition, 7/8 is the age children start learning second languages in the UK's primary schools and thus makes the results of some interest in this educational context. We have added now this info in a footnote, page 14, footnote.

41. *L. 380. Mayor & Plunkett (CogSci, 2010) showed that the acceleration in vocabulary learning cannot be attributed to (just) differences in word frequencies in the input, if words follow a Zipfian distribution. I'd be curious to know if this hold true with a geometrical distribution of words?*

We agree this is an interesting question. Ultimately it is one that would require modelling/empirical work that goes beyond the scope of this paper to answer properly.

42. *Modelling: is there a way to add variability/noise to the models (does it make sense to do this, at all??), such that differences across conditions can be statistically compared? E.g., to see if small differences across conditions are interpretable or meaningful (e.g., Table 2, the small difference between the skewed condition and the HF condition)?*

Integrating noise is possible. Considering the Rescorla-Wagner model, which updates weights linearly, it's not even necessary to change anything within the delta rule. One can simply take the asymptotic values (or the values at the end of learning as in our case) and consider it averages coming from a gaussian distribution with a fixed SD. We would then test any possible differences with a simple t-test. However the critical question is: what this SD should be? It is inevitably an arbitrary value given that we don't have a strong hypothesis about how much noise there is in the learning rate (e.g., 10% over the signal? 50%?) and why there should be noise in learning. We are reluctant to artificially manipulate noise without a clear hypothesis.

Critically, it seems to us that where there are only small differences between the models, this is the result of the simplicity of the paradigm and the materials used. Accordingly, trying to tease apart these differences would require different input sets and a different paradigm and we feel this goes beyond the scope of the project, where the purpose of the modelling is to make predictions about the learning of these input sets.

43. *Typos:*

Highlights: "spacial" should read "spatial"

L. 481: "the the" should be "that the"

L. 1174 "not meet" should read "not meeting"

Thank you for spotting these typos. We have now corrected them (hopefully now the manuscript is typos-free).

44. *Abstract: “unattested”. Should this be “untested”?*

We prefer to use “unattested” as it is the word that is typically used to refer to linguistic utterances not occurring in the input set.