Review of Stage 1 Registered Report entitled **Go above and beyond: Does input variability affect children’s ability to learn spatial adpositions in a novel language?** Submitted to PCI-RR.

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.

**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”).

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”.

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).

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.

**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?
* 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?
* 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?
* 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.
* The time interval between both sessions will somehow vary. Are you going to control for this time interval? If so, how?
* A little bit more justification of the age range would be nice. Why 7-year olds?
* 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?
* 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)?

**Typos**:

* Highlights: “spacial” should read “spatial”
* Abstract: “unattested”. Should this be “untested”?
* L. 481: “the the” should be “that the”
* L. 1174 “not meet” should read “not meeting”