We thank the reviewers for their careful consideration of our work. We first address some common concerns that the reviewers raised, and then respond to some reviewers individually.

Synthetic vs. real experiments. R1 and R4 questioned how well our analysis for the synthetic experiments in Section 3 would generalize to real-world settings. R2 suggested that an analysis on non-toy models would be interesting to see. R3 believed that the synthetic experiment was not suited to the model class. The synthetic evaluations provide three key demonstrations: 1) we introduce an "ME score" for evaluating models with respect to this bias, 2) simple neural nets have a strong anti-ME bias in common machine learning settings, 3) and the ME score decreases to zero over the course of training. Although we use simpler versions of commonly used models, the training procedures and objectives are similar to their large-scale versions. We expect our analysis on smaller models to extrapolate to larger ones (R2). In fact, we show this in Section 4.2 where we see similar anti-ME behavior of a ResNet model in a large-scale classification setting (ImageNet). We would fully expect these results to generalize to ever more complex models too where training uses a maximum likelihood objective.

Previous work. R1 and R3 enquired why previous reports of ME were not pursued and evaluated in the setups proposed in the paper. We regret that we were not clearer about how our aim differs from these studies [McMurray et al. (2012), Santoro et al. (2016), Zinszer (2018), Cohn-Gordon (2019), Lake (2019)]. As we mention in the discussion of our paper, the demonstration of ME in these models is observed only when it is explicitly built in or trained on tasks that are similar to Markman's behavioral paradigm. These setups are not designed for (and do not generalize to) the naturalistic, large-scale image classification or translation tasks considered here. It is difficult to see how these demonstrations of ME would aid downstream learning as we propose or as is observed in humans in lifelong learning settings. We will amend our discussion to clarify this point. R1 pointed us to Zinszer (2018), which is relevant to our work and we shall include the reference in our revised manuscript.

Other datasets. Both R1 and R4 suggest we analyze other datasets. R1 suggests the CHILDES corpus, and R2 suggests analysing real-world unlabelled data. These are interesting ideas, but we aren't sure how ME could be analyzed in either case, as ME requires a word learning task. We would appreciate clarification from R1 and R2 in their final review about how we could do this.

Response to R1. We thank R1 for their helpful suggestions on clarifying our use of the word "bias." In our revisions, we will also make sure to clarify which models and methods are used in each section.

Response to R4. R4 queried why we would expect a model to assign a probability of 1 to the new (correct) class. In our analyses, we wish to compare the probability mass assigned to the familiar classes relative to the unfamiliar ones for a sample from an unfamiliar class. We would like to clarify that it's fine (and even desirable) for the model to spread its mass over all the novel classes (as opposed to just the correct one). Notably, the ME measure we used is able to account for this.

R4 suggested that in the continual learning setup we train until the model overfits, and report results on a held-out set. We set up the experiment in Section 4.2 to mimic naturalistic life-long settings where some classes are seen more frequently than others. While this could result in overfitting, we see this as an inherent challenge to the learning setting—we aren't sure how a validation set could be used in that context. Our analysis focuses on examining the ME score (and the classification probabilities) on classes and samples that have never been seen by the model until that point in training.

R4 questioned the importance of the ME bias, pointing to synonymy and polysemy (where the one-to-one assumption fails to hold). We agree with R4 that ME is not universally applicable; in fact, children must also learn to overcome the bias in appropriate circumstances. Like any inductive bias, it should be a good a priori guess, but it is never the final word once data is observed. However, we demonstrate in our work that the status-quo (where models have a strong preference for the familiar classes) is highly suboptimal. Ideally, a model could autonomously decide when to use ME based on its past experience and stage of learning. Such a model should be able to accommodate for synonymy and polysemy flexibly based on evidence. For example, Lewis et al. (2020) provide evidence that the use of ME in humans is strengthened with experience. Moreover, the importance of the ME bias is highlighted in Lake, Linzen, & Baroni (2019) where the authors find ME to be central to humans learning compositional instructions with a few examples.

R4 suggested that we look at approaches where a bias is added to new classes (either to the weight or the softmax activations). We would like to point out that it is non-trivial to identify a new class. Our experiments with an oracle that tells the model whenever a new class is encountered, implements exactly what R4 suggests (we add a bias to the new class activations). The challenge here is to identify which data point would correspond to a new class. We thank R4 for pointing out two additional references that we missed. We will include these references in our revisions.

Summary. There's a huge disparity between ME in human language acquisition and the anti-ME effects that we observed in popular DNNs and MLE-based approaches. We hope that our work inspires future advances, as well as larger efforts towards incorporating biases from human language acquisition into AI algorithms. We thank you for considering our paper in your further discussions.