Dear Dr. Brown-Schmidt

Please accept the revised version of our article “The growth of children’s semantic and phonological networks: insight from 10 languages” (COGNIT-D-18-00503). In this second round of revisions, we developed the general discussion to accommodate the suggestions made by R1 and R3. We also explained why the conclusions of this study are valid without having to run the additional experiments suggested by R3. We highlighted in yellow the parts of the new manuscript where changes were made. We hope that this revision proves suitable for publication. Please do not hesitate to contact us if you have any questions or concerns.

Sincerely,

Abdellah Fourtassi & Michael C. Frank

Editor's Comments

> *R1 and R3 offer suggestions for clarifying the framing and interpretation of your findings that I encourage you to consider.*

We enriched the discussion to accommodate the suggestions made by R1 and R2.

> *R2 raises several concerns, among them that the paper does not allow characterization of individual performance for the manipulations explored in Experiments 2-3. Given that Experiment 4 accomplishes this by adding many more measurements per person, it may be worthwhile discussing why a similar approach wasn't (or doesn't need to be taken) for the manipulations of Experiments 2-3.*

We provided, below, our argument explaining why running additional experiments is not necessary to show suboptimality. The argument can be summarized as follows: As much as the claim of (near-)optimality should be verified at the individual level (since it is possible to have a spurious optimal pattern on average while individuals show suboptimal behavior), it is not necessary to verify sub-optimality for every individual – since a sub-optimal average pattern cannot emerge from a predominantly optimal behavior at the individual level.

Reviewer 1

> *First, with regards to the theme of the paper, while I do see the link to the task of word learning at a computational level, I do not think all readers would share this conception. Like other reviewers have pointed out, there are some logical leaps to directly this work to (child) word learning. In the reply to my first round of review concerning the use of namable objects, the authors said that this is analogical to second language learning and mapping new words to existing concepts. While I am not sure this can fully resolve all the potential questions of the readers, I think explicitly making arguments as such in the introduction (or discussion) can better motivate the link of the current experiment to (child) word learning.*

We added the following two paragraphs in the discussion. The first explains why we used familiar categories instead of novel ones and the second argues that the fact of using familiar objects is still relevant from a word-learning point of view.

"Note, however, that though our framework allows in principle for the study of how sound-meaning interaction influences the underlying categories, we did not explore this question in the current task. In fact, we used familiar categories both at the semantic level (cat and dog) and at the phonological level (the phonemes /p/ and /b/). This choice allowed us to test

the mechanisms of cue integration separately from the question of category learning. Using novel categories would create a confound as participant would additionally have to learn the boundaries of the novel categories, making it difficult for us to tease apart how much of the suboptimal behavior is due to the difficulty of audio-visual integration and how much it is due to the difficulty of estimating the boundaries of the novel categories."

"Though our task deals primarily with the mechanism of word recognition, the conclusions of the current study could potentially apply to some word learning situations as well. While the typical task in the word learning literature involves mapping a new label to a novel object (e.g., Markman, 1991), several cases of word learning in the wild involve familiar objects (as in our task). For example, adults learning a second language do map new labels to objects that are already familiar to them. Besides, children come to the task of word learning with fairly refined knowledge about semantic categories, at least in the case of basic-level object categories (e.g., Quinn et al., 1993). That is, in many situations, they do not associate new labels with objects that they are seeing for the first time (or for that matter, labels that they are hearing for the first time). Rather, they often map familiar sequences of phonemes onto concepts that they have refined over the course of earlier exposures (e.g., Bloom, 2000)."

> *With regards to Figure 5, I notice that the subjects' actually performance was systematically closer to chance than the model's prediction. Could it be caused by subjects sometimes failing to pay attention to the stimuli and resort to random guess? If that is the case, and if simply adding a lapse rate can increase the fit of the optimal model to be comparable to the descriptive model, this can provide more evidence towards the discussion as to whether the combination of two cues is optimal. Whether lapse rate can be estimated of course depends on the amount of data available to fit the model (per subject) and the variability across subjects, so I am not sure such a model will actually work. I'm proposing this more out of my own curiosity.*

This is a good observation. We address this issue in the paper:

*"Figure 5 shows that the participants deviated slightly --- but systematically--- from the optimal prediction in that they were slightly pulled toward chance (i.e., the probability 0.5). This fact was captured by the increase in the value of the variance associated with each modality (as can be seen in Table 1). Note, however, that despite this increase in response randomness, our analysis of modality preference showed that the *relative* values of these variances were not different (Figure 5), meaning that there was no evidence for a modality preference."*

> *In other figures and tables, Auditory comes before Visual but in Fig 5 it's ordered different. Keeping the ordering consistent will help the reader follow the paper better*

Thanks for catching this. We rearranged the order of "Visual" and "Auditory" in Fig 5.

Reviewer 2

> *...The inclusion of this experiment only partly addresses my concern. While it does suggest that the population-level findings from Experiment 1 are also seen at the individual level, this experiment does not speak to the results from Experiments 2 and 3. The concerns I raised regarding the analysis approach are also applicable to the conclusions that can be drawn from these experiments. Indeed, given that these experiments involve adding noise to one of the modalities, the results from these experiments are even more susceptible to the possibility that individual participants might differ in the baseline extent to which they are subject to uncertainty in each modality.*

We stated in the manuscript that the goal of Exp 1-3 was to study behavior at the population level, that is, we were interested in the global, average tendency, not in the characteristics of individual behavior (a similar approach was taken in previous studies such as Feldman, Griffith & Morgan, 2009; Kleinschmidt & Jaeger, 2015). We agree that the claim of optimality made from Exp 1 can be particularly problematic: It is possible to have an optimal pattern on average which spuriously emerges from the aggregation of suboptimal individual responses. This is the reason why adding Experiment 4 was crucial to rule out this scenario. Nevertheless, the converse is not true: It is not possible that a sub-optimal average pattern emerges from the aggregation of predominantly optimal behavior at the individual level. In other words, though *some* individuals may be optimal in Exp2 and Exp3 (where we introduced additional noise), they cannot represent the typical behavior in the group (otherwise the average would be optimal, as we found in Exp 1).

> *...As such, the authors haven't actually addressed my original concern, and indeed the removal of these estimates makes it harder for readers to get a handle on how noisy the sample under consideration is. If that approach was inappropriate, is there an alternative approach they can take to provide the reader with some idea about the extent to which individual participants' behavior deviated from the population-average? And does this approach show that there isn't a high degree of variability in the sample? As I argued in my previous review, and related to point 1 above, the use of population-level analyses is particularly problematic when there is a high degree of variability in the sample.*

As we said in our previous response, the goal was to study behavior at the population level. The fact that there may be individual variability does not speak against the fact that the aggregate captures the global tendency, especially in the case of suboptimality (as reported in Exp 2 and 3).

Reviewer 3

The first was addressed by the authors by suggesting that their task is different enough from Sloutsky & Napolitano's (2003) and Napolitano & Sloutsky's (2004) papers in that, a) the instructions were that these were supposed to be words, and b) that their task encouraged categorization, while those previous papers somehow did not. I find the first line of argumentation to be somewhat weak (i.e., is this meant to suggest that this simple linguistic instruction would lead to the use of different cognitive mechanisms of little interest to the auditory-dominance literature?).

We added the following paragraph to discuss how our results relate to that literature:

"Our task was an adaptation of a task introduced originally by Sloutsky and Napolitano (2003). This original work is part of a rich literature about modality-dominance in cross-modal processing (Robinson and Sloutsky, 2010, Colavita, 1974). Generally speaking, this literature suggests that, when presented with an audio-visual stimulus, adults pay more attention to the visual component whereas children may pay more attention to the auditory component (Hirst et al.,2018, Barnhart et al., 2018). Our work shows that adults do not systematically prefer one modality over the other. Rather, they modulate this preference as a function of noise (Experiments 2 and 3). That said, it is important to keep in mind that while the literature on modality-dominance deals with low-level perceptual encoding, our task targets specifically high-level linguistic categorization. It is possible that the underlying mechanisms in both cases are related (or even similar). However, as it stands, our work cannot speak to this possibility."

The second line of argument, that S&N/N&S's task did not involve categorization, also seems erroneous. At least in N&S (2004), children were taught a general property of the stimulus model (i.e., the first target presentation, equivalent to the endpoint stimulus here): That it would reveal a prize. Then, on the test items, children were again asked to choose whether it might lead to a prize. This looks a lot like categorization, and thus seems to be very, very similar to the current task (invalidating the authors' second line of argument).

Our task was inspired by SN, not NS. Thus, our response was based on the comparison with that same work which did measure perceptual discrimination, rather than linguistic categorization. The task introduced in NS is interesting and can be understood as involving some categorization, but the degree to which such categorization maps on to linguistic processing is unclear.

Let me just conclude by saying that there is indeed a rich literature about developmental modality-dominance and intersensory perception that is directly related to these exciting findings. It is my opinion that the authors are doing a disservice to themselves by not making any attempt to discuss how their results are relevant to that literature. For example, if the authors move forward to test these results in developmental populations, as they argue in the discussion, the work of Robinson (Barnhart, Rivera, Robinson, 2018, for a recent example) as well as the large literature on the Colavita effect (e.g.,

https://www.sciencedirect.com/science/article/pii/S0149763417307674) would make some concrete predictions about this. Currently, none of these findings are discussed. I do ask that the authors reconsider this consideration moving forward.

Thanks, we appreciate this comment. See the paragraph above where we discuss the results in relation to the literature on modality-dominance in cross-modal processing.

p. 9, lines 173-175, the text does not match the figure (in that the example given in the text is not the same example in the figure). Potentially confusing for readers.

We changed the text to be consistent with the figure.

p. 15, p. 38, lines 687-693, the paragraph gives the false impression that children's lexicons have particularly dense phonological neighbourhoods when this is in fact incorrect (e.g., Coady & Aslin's work). I suggest minor rewording here.

We appreciate the comment and have revised the paragraph slightly.