Action Editor

> *I ask that a revision center on clarifying the state of knowledge on children's distribution of eye gaze in language comprehension to highlight the unique features of this work (there are many, but bringing those to the foreground would be useful to readers unfamiliar with this area of study).*

Thank you for pointing out the need to clarify the prior theorizing about children's eye gaze during language comprehension. We added text to the introduction and to the section on vision-language interactions, which we think provides a clearer overview of the current understanding of language-driven eye movements and highlights the unique contribution of our proposal – that eye movements reflect more than the output of the language comprehension process and are fundamentally linked to listeners' internal states (goals/knowledge).

> *You should also work at ensuring that all analyses are clearly explained and connected to behavioral data; make consistent your references to the studies (I personally agree that the term "case study" does not work); and consider whether there is a way to streamline results to spotlight take-home messages.*

Thank you for this point. Following Reviewer #2's suggestion, we have updated the text to focus on framing E2 as an experimental follow-up to the more exploratory comparison in E1, allowing us to move away from the "case study" framing.

> *Also, I agree with Reviewer #4's comments around reaction times, and the need to think carefully about their onsets.*

We appreciate the point about how to best operationalize reaction time (RT) in this task. We think that there is a good reason based on prior literature to define RT as the latency to shift gaze away from the central stimulus to either object measured from the onset of the target noun. Each target word/sign pairing was selected such that there would be minimal phonological overlap at the onset (e.g., cat-bird). We have added text to the analytic approach to justify this decision.

> *This reviewer also requested, aptly, a clearer motivation for language-driven and non-language-driven shifts, and a rationale for the inclusion of adults in Study 2.*

We added text to the introduction of Study 2 that motivates the inclusion of the adult sample by hypothesizing a potential developmental difference such that children would rely more on visual information in the noisy context because they have less-developed perceptual models (following Yurovsky, Case, & Frank, 2017). We also speculate that we might not have detected a difference in the current paradigm because we used such highly familiar target words.

> *I concur with the suggestion that additional non-Bayesian analyses be reported for clarity (as supplemental data if you deem appropriate).*

Thank you for this suggestion. We added a parallel set of non-Bayesian analyses to the appendix.

> *You might want to take advantage of a new policy as suggested in the editor's interview on the journal's web page: "I also encourage authors to include a brief paragraph at the end of the article describing its broader context, explaining such things as how the ideas originated, how the findings are related to the authors' research program, and how the research will be extended in the future, much as one provides in oral presentations." You may label it something like Context of the Research if you wish. By brief, I mean about the same length as an abstract, not more than about 250 words.*

Thank you for this suggestion. We added this section.

Reviewer #2:

> *Although I found the results convincing and consistent with the proposal, I also found myself wondering: what (if any) alternative accounts there are about how children deploy gaze in language comprehension, as well as how this account is different from how researchers currently think about children's gaze behaviors during language comprehension. I think devoting some space in the Introduction for the authors to state the current state of thinking (or the lack thereof) about how and why children distribute gaze in language comprehension would further clarify the empirical contributions of the current research and strengthen the theoretical impact of this paper (and of the author's proposal).*

Thank you for pointing out the need to better situate our theoretical account, especially with respect to alternative proposals. We added text to the introduction and to the section on language-vision interactions that serve this goal. Specifically, we highlight how the current work on children's distribution of attention during language comprehension leaves open the extent to which eye movements are driven by listener-external (salience) vs. listener-internal (goals/knowledge) processes, and whether this might change across development as children build better language and perceptual models.

> *...the link between the authors' current research and the visual-world paradigm was not as clear to me (beyond the methodological similarity that both use eye-tracking to study language comprehension).*

This is a good point – thank you. We added text to VWP section to clarify the theoretical connection with our work. We highlight how prior studies have focused on eye movements as a measure of language comprehension, whereas, we propose that it is important to include eye movements that are driven by the goal of understanding language when considering visual attention during language comprehension.

> *Also, I found the literature reviewed in the discussion section on "adaptation of visual processes" possibly relevant to help the authors frame their proposal in the introduction. I thought I would share these comments in the case the authors find themselves revisiting the structure of their introduction.*

Thank you. We now mention the idea of adaptation to different experiences in our discussion of ASL in the introduction.

> *That is, I was able to follow what "boundary separation" and "drift rate" parameters reflected. However, I was unable to grasp how these two values would get manifested in the behavioral data. That is, is it that slower language-driven RTs would count towards greater boundary separation and faster language-driven RTs would count towards greater drift rate? The extent to which the authors state this would help clarify the modeling portion of this work.*

Thank for this point. We updated our description of the models to include a clearer description of how changes in parameters map onto changes in behavior: i.e., slower but more accurate RTs would result in a higher boundary separation estimate, while faster and more accurate RTs would result in a higher drift rate estimate.

> *Related to the modeling piece, how was the pre-defined upper control limit (in the EWMA) and the pre-defined decision threshold (in the DDM) set and how did this impact the outcomes of the model (e.g., the relative contributions of boundary separation and drift rate estimates)?*

We set the width of the upper control limit to be two standard deviations away from a model of random responding where the probability of a correct shift equal to 0.5. We did not systematically vary the value of this parameter since it was fixed across the groups/conditions, so it should not influence our comparison of the model estimates. We updated the EWMA model description to include this information.

For the DDM, we estimated the decision threshold for each condition from the data. Since the terminology is quite similar across the two models, we have added text to the analytic approach section to help make it clearer which parameters were set beforehand and which were estimated from the data.

> *At different points in the manuscript, the nature of the two experiments are presented slightly differently. In the abstract and the intro, the two experiments are conceptualized as "two case studies" testing a single proposal. In the transition from Experiment 1 to Experiment 2 (and the discussion), Experiment 2 is conceptualized as a well-controlled*

> *follow-up to the more "exploratory" and "post-hoc" analysis of Experiment 1's data. Although these two conceptualizations are clearly not mutually exclusive, I did think the paper would benefit from a more consistent conceptualization of the pair of experiments throughout the paper. I also would recommend favoring the second conceptualization (i.e., that Experiment 2 is an experimental follow-up to the observational Experiment 1) since I believe it makes the two experiments a more compelling package.*

Thank you for this suggestion. We updated the text to focus on the framing of E2 as an experimental follow-up, moving away from the “case study” framing.

> *Given the data-rich nature of this paper, I thought the authors might want to consider paring down the analyses / conditions in this paper in order to let the key comparisons shine more. In particular, the analysis of the bulls-eye and object conditions in Experiment 1 seem secondary to the main comparison of ASL to spoken English stimuli. Limiting Experiment 1 to only those two conditions might help focus the readers' attention on the key findings.*

We appreciate the goal of streamlining the manuscript to highlight key results. We do think that the Bullseye and Object conditions are relevant to our account. To better communicate the value of these conditions, we added text to the Discussion of E1 that better highlights the contribution – that is, within spoken language learners, there is evidence that an informative location (speaker’s face) leads to slower and more language-consistent gaze shifts.

> *Minor comment: the description of the ASL linguistic stimuli (specifically the sentence structure) in the text does not match the description in Figure 1.*

Fixed. Thank you.

Reviewer #3

> *Introduction: I would suggest that the authors cite Henderson TICS 'Gaze control as prediction' and discuss the predictive element of deciding when/where to look. I think the results are consistent with Hayhoe and Ballard's claim about information seeking and reducing uncertainty, but Henderson's take is a bit different...it emphasizes prediction about what spatiotemporal locations will be informative and there's certainly an element of that in the current study.*

Thank you for pointing us to Henderson’s interesting TICS paper. We added a discussion about prediction-based gaze control to the section on goal-based accounts of vision in the introduction.

> *Introduction: The authors wrote: "Across both experiments, the critical behavioral prediction is that listeners will adapt the timing of their eye movements to facilitate*

language processing". Typo with 'language'. I also think this statement could be more specific. There is evidence of adaptation of the timing and target of eye movements; timing alone isn't too helpful if you don't pick the 'right' place to look.

We fixed the typo – thank you. We also updated the sentence to mention “informative locations” to reflect the temporal and spatial nature of the prediction.

Analytic approach: I appreciate this section to explain each of the analytic methods used. I think that it is all clear except for the HDDM section. I found myself flipping back to this throughout the paper but never felt like I got the intuitive grasp I needed about what boundary separation and drift rate means for the particular data analyzed.

Thank you for pointing out that we could have been clearer in linking the HDDM to the behavioral data. We updated the following text (new parts in bold) to the analytic approach section that describes the behavioral pattern that changes in boundary and drift would capture.

“Here, we focus on two parameters of interest for our hypotheses: *boundary separation*, which indexes the amount of evidence gathered before generating a response (**higher values reflect slower but more accurate first shifts**, suggesting more cautious responding) and *drift rate*, which indexes the amount of evidence accumulated per unit time (**higher values suggest faster and more accurate first shifts**, suggesting more efficient processing).”

Experiment 2 method: I think more details are needed about the eye tracking data. Was calibration accuracy assessed and did it vary by age? I don't think it's likely to be a problem but would be good to rule out that poor accuracy could relate to the observed effects.

Thank you for this suggestion. We added the following information about calibration accuracy to the methods section:

The average calibration accuracy for calibration was 0.39 degrees of visual angle, with similar values for children (M= 0.38, SD = 0.21) and adults (M= 0.39, SD = 0.14), and little evidence for an effect of age within the child sample (βage = -0.03, [-0.09, 0.03]).

Experiment 2 method: How were data excluded based on missing eye tracking data? The preregistration document on OSF says 50% of data loss would lead to trial exclusion, but it seems like in this paradigm one would need to be more strict to collect intact timeseries for these analyses. It would be helpful to explain how this was done, how many trials were contributed by participants, and whether this depended on age and condition.

Thank you for pointing out that we needed to provide more information about the data exclusions. We added more details about the average number of trials contributed for each condition in both Experiment 1 (see Table 1) and Experiment 2 (see the Participants section). In Experiment 1, ASL learners contributed fewer trials on average likely because of the younger ages included in this sample. In Experiment 2, we followed the 50% trial and subject-level exclusionary criterion, and after applying these exclusions, adults contributed on average 30.5 trials (min = 17, max = 32) and children contributed on average 26.8 trials (min = 14, max = 32). There was no difference in the average number of trials contributed for the noisy (14.3) and clear conditions (14.2). We think the methods section is much improved based on this suggestion – thank you.

> *Figure 4: It is difficult to discern center and target in child panel because they cross at the last moment.*

Thank you. We moved the labels for the children's target looking curves.

Reviewer #4

> *Reaction time was calculated as delay to look to target following onset of noun, but the stimuli are different in length (not only the word length, but it also seems like length of entire trial). The difference seen between ASL and English groups could be due to the word not being completed yet, as the crossover point in Figure 2 for looking at target versus center seems to be relatively similar for the ASL and Face groups, when considering the offset of the word.*
>
> *a) The authors should repeat their analyses, but with their definition of reaction time taking into account the offset of a word (not just its onset).*
>
> *b) The authors should also look to see how many "language-driven shifts" occurred before versus after offset of the word.*

Thank you for these suggestions. First, we want to point out that it is a very strong convention to calculate reaction time in the visual world paradigm from word onset rather than offset. This convention dates at least to Allopena, Magnusson, & Tanenhaus (1998) in adults, and to Fernald et al. (1998) in children. The reason for this convention is that incremental processing of word forms begins at the onset of the first sound. Both adults and young children (significantly younger than those in our sample) show this kind of incremental processing across a wide range of experiments. Thus, aligning reaction time to word offsets in general fails to "give credit" for the processing that begins at word onset and adds variability to estimates of reaction time because of variation in word lengths.

In our dataset, however, the consequences of alignment to offsets would complicate interpretation even more. Words and signs were different lengths across ASL and English stimulus sets, thus alignment to offsets creates systematic group-level biases in reaction time. We also want to provide a few other comments regarding the issue of onsets/offsets:

- All of the words and signs in our experiment had minimal phonological overlap at noun onset, meaning that from the very first moment in the signal, they could be differentiated.
- In our prior work (MacDonald et al., 2018), we specifically analyzed the probability with which children and adult ASL users responded during vs. prior to sign offset. As illustrated in Figures 2b and 2d in that paper, the mean proportion sign length processed prior to a shift in visual attention was less than 1.0 for both adults and children. While there was some small variation as a function of sign length, this pattern held for all signs processed by adults and for all but 1 sign for children. These analyses suggest that both children and adults generate rapid gaze shifts prior to sign offset during ASL comprehension. This further suggests an account where ASL-learners, like children learning spoken languages, process signs incrementally and do not wait until the end of the sign to seek a named object. We have added text to the current manuscript pointing readers to these analyses and their interpretation relevant to the current set of studies.
- As also mentioned in the point below, the length of signs did not vary across condition; the words were the same length across the Face, Object, Bullseye, Clear, and Noise conditions. We found evidence of slower but more accurate shifts in the Face context relative to the other two conditions and in the Noisy relative to the Clear context, speed-accuracy tradeoffs that could not be related to varying word lengths.

Taken together, we think that the interpretative concerns related to E1 are best addressed by having a well-controlled, experimental follow-up study. Thus, we chose not to conduct additional analyses of E1 since we were not sure they would sufficiently address the concerns about interpretation and we added text to the limitations section of E1 to make this point clearer. We also added text to the Visual World Paradigm section to clarify that prior work has measured RTs from word onset.

> *The distinction between language-driven and nonlanguage-driven shifts needs to be better motivated and defined. There seem to be a lot of assumptions in the categorization process that need to be addressed, especially as the subsequent drift-diffusion model builds on these analyses.*

Thank you for this suggestion. We updated the text to clarify that we are only using the terms language-driven and nonlanguage-driven as verbal labels for what the EWMA model measures. That is, language-driven just refers to the proportion of shifts that deviate from a model of random responding on a 2-AFC task with a predefined upper limit on what would be considered guessing behavior. This is a label that is only relevant within the small world of our analysis model, since the nonlanguage-driven shifts could be motivated by lots of goals, including language-related goals. We appreciate this point, and thus, we chose to change the label of “nonlanguage-driven shifts” to “random shifts” to better reflect what we think the EWMA is measuring.

For the DDM analysis, we only model language-driven shifts since we want to be confident that these responders were generated by the underlying process of interest. The DDM was not designed to incorporate responses generated by non-decision-related processes such as those made prior to accumulating enough information to respond (i.e., guessing). In our context, the EWMA acts as a filter to provide better estimates of the decision-processes that generate eye movements, removing contamination from fast guesses.

> *In both experiments, the authors should test to see if length of word has any effect on reaction time (as this was variable across groups and conditions).*

Thank you for this suggestion. As discussed above, previous analyses as reported in MacDonald et al., (2018) have partially addressed this issue by suggesting that both adult and child learners of ASL process signs incrementally, starting from the onset of the sign, as has previously been shown in children and adults processing spoken language. This was true regardless of the length of the sign.

With respect to our experiments, the same word length was used in the Bullseye/Object/Face conditions in E1 and Noisy/Clear conditions in E2 (for all groups), so while there might be some variability attributable to sign length within that particular condition in E1, it should not be a factor influencing our estimate of differences between spoken language conditions.

> *While I applaud the authors for their dedicated use of Bayesian statistics, I strongly recommend the addition of frequentist tests as well (whether in the main article or as supplementary material). This would provide added transparency for the average reader unfamiliar with Bayesian stats.*

We added a parallel set of frequentist models to the appendix.

> *In the Behavioral Analyses section of Exp 1, the authors compare curves without sufficient statistical tests. There were analyses done on the curves for looks to the center stimulus, but not the target and distractor.*

We chose not to analyze the target or distractor looking curves because the critical prediction was a difference in the time course of looking to the social target (center fixation). We have made this explicit in the results section.

> *The authors argue for a difference in looking to the speaker between the Noisy and Clear conditions based on "visual inspection" of Figure 4a (page 27). My own visual inspection does not suggest any substantial difference. Though there is significant difference, the authors are encouraged to provide summary statistics for these two conditions as well as some measure of effect size.*

Thank you for this comment. We have made it clearer in the text that we use the term “visual inspection” simply to point the reader to the rightward shift in the center-looking curves in the Noise condition, “suggesting” an effect of the manipulation. Moreover, we more clearly point out that the interpretation is statistically supported by the cluster-based permutation analysis. We’d like to report an effect size here but don’t know how to do so in an unbiased way. Using the classic approach of taking the average difference between conditions in some window, the size of the proportion looking difference would depend on the time window chosen, which we did not specify ahead of time. Thus any effect size estimate would be post-hoc. Given this situation, we chose not to provide an effect size.

> *The reaction time analyses in Exp 2 are unclear and haphazard. The authors report mean differences to identify the target between the two conditions for adults and children but appear to test for significance collapsing across the two samples. The differences in RT should be tested separately in adults and children.*

We added the appropriate condition contrasts for RT within each age group.

> *It is reported that older children responded faster than younger children, a similar test should be done between adults and children.*

Thank you. We added the contrast between adults and children.

> *No statistical tests were done comparing the HDDM parameters for adults and children in Exp 2, though differences between the two groups are described.*

Added.

> *There is no statistical test on the drift rate parameters for Exp 2.*

Added.

> *The authors do not sufficiently motivate their use of adults in Experiment 2. A clearly stated hypothesis about developmental effects they expected to see in their study is also needed, as it is hard to infer from the citation on page 10.*

This is a good point. We included adults because we thought that children might show an even larger difference in looking behavior in the Noise condition since they have less exposure to the word-object mappings as compared to adults. We added this hypothesis to the introduction of Experiment 2, and we also added some interpretation of this result in our discussion of Experiment 2.

> *Are there any reasons to suspect differences between deaf and non-deaf ASL learners? This should be addressed. The non-deaf children may have learned to use social cues differently based on their exposure to spoken language.*

The hearing status of ASL learners might be an important factor that we have addressed in a previous analyses in MacDonald et al., (2018). In those analyses, there were few differences in the looking-time behavior of deaf vs. hearing ASL learners. In that paper and in the general discussion of the current manuscript, we speculate about how future research is needed to understand whether the differences we observed in E1 are driven by a change in a general response strategy or by the in-the-moment constraints of processing a visual-manual language in real-time.

> *It is recommended that the color of the grey lines in Figure 3a be changed to a more visible color.*

Thank you for this suggestion. We switched to a darker color and increased the line size to make the curve more visible.