Dear Dr. Ganea,

Thank you for the thoughtful comments on our manuscript, “Real-time lexical comprehension in young children learning American Sign Language.” In our resubmission, we have addressed your comments and the comments of the reviewers, as described in the point-by-point response below

A point of concern for both reviewers was the theoretical framing of this work. To address these concerns, we would like to highlight the updated structure of our Introduction and Results. Following Reviewer 2’s recommendations we now first present the evidence that children and adults generate rapid gaze shifts prior to sign offset during ASL comprehension. Then, we present the group comparisons (children/adults and hearing/deaf) followed by evidence of links between processing efficiency and productive language development. With thanks for these suggestions, we think that this new structure, along with the expanded introduction and discussion, focuses the paper more clearly on a streamlined set of issues and clarifies the motivation for our group comparisons.

Please let us know if you have any questions or concerns. We look forward to your consideration of this revision.

Sincerely,

(author name removed for blinded review)

Reviewer 1

> *I think I just have some initial confusion about vagueness of terms such as "modality" and "visual attention". I understand the introduction starts with the big picture, yet some terms still need defining up front. For example, "processing efficiency" is mentioned on the first page, and it is not yet defined, and I have no idea what it refers to yet. I mention below some areas where the introduction needs more motivation (linking ideas). Also, I'm not clear on the mechanisms driving the predictions.*

Thank you for pointing out places in our manuscript where we could use clearer language. Please see below for a point-by-point response to each of your suggestions.

> *Moreover, the discussion feels rather short and amiss -- what are the implications for education, deaf children? Are they able they speak to the issue about ASL being is a natural language that is processed in the same way as spoken languages are? And that their study suggests common neural architectures for language processing and that ASL is not "bad" for the deaf child's brain?*

Thank you. We expanded our discussion and added some thoughts about the implications of our results for deaf children's learning. Specifically, we point out that, similar to spoken language development, early ASL processing efficiency is one link in a chain that starts from the quality of children's input, which leads to better vocabulary development, which in turn results in higher levels of academic achievement. However, we hesitate to make any claims about common neural architectures or the value of ASL exposure since these were not the focus of the current work.

> *1. First sentence of abstract is confusing because if children are interpreting spoken language, yet linguistic information drives rapid shifts in visual attention, then what are they looking at if they're interpreting spoken language.*

Thank you. We have clarified the first sentence.

> *2. And the last sentence of the abstract is confusing too because at this point it is so very vague to refer to "visual attention", at least inform the reader if you mean visual attention to look at a signer or to look at possible referents in the world (that may or may not be present).*

Thank you. We have clarified the last sentence in the abstract.

> *3. The first paragraph of the introduction is a little vague, and I'm not sure how things link up – for example, the last sentence in this paragraph on line 34 – what does this have to do with eye movements, mentioned in the previous sentence?*

Thank you. We have clarified this sentence.

> *4. Page 3, I'm not sure what you mean by "modality", given that this study is entirely in the visual modality: What do you mean by, "Do the findings from spoken language reflect language-general phenomena or are they specific to the auditory modality?" I'm guessing you are referring to looking at the referent before the full word is uttered, but that's a pretty superficial prediction about how processing speech and sign differ, given that the time course of speech and signs differ, besides modality. Basically, I just think that question isn't clear yet.*

We added citations to the specific findings of interest in the literature on spoken language development -- that is, the relations between variability in processing speed as indexed by the timing and accuracy of eye movements and children's productive vocabulary.

> *5. On p. 6 lines 24-36 there is a nice prediction (the second out of two) - possibly ASL learners, like spoken language users, would shift visual attention as soon as there's enough linguistic information. But I feel this begs qualification about basic human vision - is it possible that ASL learners could shift attention to the referent but nonetheless still perceive the signing in the parafovea(?) ... that's certainly plausible so that's a third prediction (or at least a qualification about the second prediction) and should be mentioned... This also comes up on pg. 23 line 52-55 – maybe children can still see the sign in the parafovea while looking at the target object.*

Thank you for pointing this out. We have clarified that shifting away from the signer reduces the quality of the linguistic signal, even though they could still be perceiving some information in the parafovea. (pg. 4 and pg. 28)

> *6. Page 6 line 24 - why are ASL LEARNERS compared to spoken language USERS? You've set up the hypothesis to sound like you are comparing ASL processing vs spoken language processing, to see if both show incremental patterns. I just wanted to point out that oddness, because that's not what you are doing.*

Thank you. We have gone through and made sure that our terminology is consistent throughout.

> *7. The 2nd research question focuses on age-related differences, so I'd like some review of the literature on how age influences visual attention in hearing kids vs adults. The 4th research question focuses on "expressive vocabulary development" as well as age -- So what's the literature on how vocabulary size influences visual attention?*

We added citations to work on the development of visual attention throughout later childhood. The upshot of this work is that different components of visual attention (e.g., the ability to distribute attention across the visual field, attentional recovery from distraction, and multiple object attention) develop at different rates (Dye and Bavelier, 2009). We also included discussion of work on children's developing ability to disengage from a central stimulus to attend to stimuli in the periphery between the ages 7 months and 14 months (Elsabbagh et al., 2013).

We also expanded our discussion to include the point that factors other than speed of lexical access (e.g., increases in general processing speed and/or control of visual attention) could lead to improved performance on the VLP task. However, there is a large body of work showing that aspects of language (e.g., frequency of words, neighborhood density, and amount of language input) affect the speed and accuracy of eye movements in the Looking-While-Listening style tasks. These language effects suggest that these eye movements are indexing lexical access as opposed to other underlying constructs.

> *8. Measure of ASL expressive vocabulary size - it says "developed specifically for this project" so clearly some further adaptations/modifications were made on top of Anderson & Reilly which itself is a modification of the CDI for ASL users. So it needs more qualification, what was changed, why wasn't Anderson & Reilly sufficient by itself? How would others replicate this?*

Thank you for pointing out the need to justify this design choice. We decided to create our own version of the ASL-CDI because we anticipated collecting data from a larger age range than the Anderson & Reilly version was designed for. To facilitate the reproducibility of our work, we will share all of the study materials (linguistic stimuli, vocabulary measure, and the gating experiment) and analysis code.

> *9. "Matched for visual salience" means what?*

Thank you. This point has been clarified.

> *10. P. 10 line 26 "chosen based on naturalness" – how?*

We have added information about how the final tokens were selected.

> *11. P. 10 line 33-36 - "minimal phonological overlap" really needs to be defined. How did each pair overlap? location, movement, handshape? Be specific. What was the*
> *definition of "minimal" - 1, 2, 3 parameters, etc.? Can they reference it with ASL-LEX?*

Thank you for referring us to ASL-LEX -- it is a nice resource. And thank you for pointing out that the phrase “minimal phonological overlap” needed to be defined. Using ASL-LEX, we can now say that our item-pairs vary in the degree of phonological overlap from 1-4 parameters. We changed the description of our stimuli in the paper (see Table 2) to be more specific about the degree of phonological overlap and the iconicity for each pair of signs.

However, because we defined the disambiguation point of each item-pair empirically, the degree phonological overlap should not have a large influence on real time lexical processing. We also added an analysis of the effect of phonological overlap and iconicity to the supplement. We did not see evidence of either factor playing a role in the timing or accuracy of eye movements in our data, but it’s important to point out that our stimuli were not designed to measure these associations.

> *12. Was the trial structure just like how it is presented for auditory LWL tasks (end of P. 10)? If not, explain the differences*

The trial structure was highly similar to that of the auditory LWL task. The only change was the addition of a two second still frame after the target sentence. We made this decision in order to to give children additional time to shift their gaze from the signer to the objects. We have updated the manuscript to make it clear that this is the only deviation from the auditory LWL task.

> *14. Computing measures of sign onset/offset by adults (explained on page 11). I would*
> *really like to see more information on this...how much variation was there among the 10 adult signers in figuring out sign onset? And really they also measured the time it took to press the button, so their definition of "sign onset" time is actually sign onset + button-pressing time.*

Thank you -- this is a good point for us to clarify. We agree that the “sign onset” measure derived from our gating experiment does not map directly onto the task of processing of ASL in real-time. It is important to point out that we did not measure response times in the gating task; instead in our gating paradigm, we measured adults’ accuracy after varying amounts of the linguistic signal. Participants had as much time as they wanted to generate a response, so time to press the button is not a component of this measure. We have clarified this in the text.

With regards to variation in this measure, there was no variation in the choice point because we selected this value when all 10 signers reached 100% agreement.

> *15. How would the addition of auditory sensory experience change the way in which CODA children process ASL vs deaf children? They're both looking at the same language, so how could the simple fact that one group can hear (and sound isn't a part of this study) and the other cannot change how they process ASL visually (they're both natively exposed to ASL). I just want this to be clearer.*

Thank you for pointing out that we need to clarify the theoretical motivation for this hypothesis. We have added text to make it clear that this hypothesis is motivated by prior work that has used comparisons between native hearing and deaf signers to dissociate the effects of learning a visual-manual language from the effects of lacking access to auditory information (e.g., Bavelier, Dye, & Hauser, 2006).

> *16. Why do adult signers show larger CI's for proportion of looking (Figure 2), compared to children? I'd expect it to be the opposite.*

The larger confidence intervals reflect that we have approximately half as many adults participants (16 adults, 29 kids).

> *17. Pg 20 Line 54 - "like children learning spoken language, ASL learners improve...over the second and third years of their life..." but this age-based development is never really fleshed out in the introduction. Ditto for pg 22 line 29*

Thank you for this point. We have added the relevant citation.

> *18. Pg 23 line 19: "While the hearing children could use vision and hearing to process incoming information, this experience did not change the timing of gaze shifts during ASL comprehension as compared to their deaf peers." - I do not accept that...they could NOT use hearing to process incoming information, there was no voiceover or audio. Why should they even expect a difference between deaf and CODA when viewing ASL? This part just needs more theoretical elaboration and motivation.*

Thank you -- see the response to point 15.

> *1. Page 5, Are you sure to say that lexical decision tasks are not "on-Line"? I am quite sure they are called online, at least the cross-modal lexical decision task with priming effects is considered a measure of real time lexical access.*

Thank you for pointing this out. This has been clarified in the text.

2. Why Figure 2 and 4 b/w but other figures are color?

Figure 1 is in color because it provides a direct representation of what participants saw in the task. But we try to use black and white because it makes our paper more accessible across the widest variety of viewing conditions and for those people who do not process color information.

3. P. 9 line 12-15 - "visual distractions" twice in one sentence

Fixed. Thank you.

4. How far away was the baby from the screen?

Thank you. The child sat on the caregiver's lap approximately 60 cm from the screen. We have added this information to the text.

Reviewer 2

> *1. One concern is that methodological and theoretical issues get a bit tangled at various points. This could be readily handled with some rewriting and reorganization. As background, there is (as the authors note) existing evidence that signs are interpreted incrementally by adults (Emmorey & Corina) and that this can be assessed using a modified listening-while-looking methodology, where participants have to shift gaze from the signer to visually co-present referents (Lieberman et al.). So the immediate question is whether the latter methodology can be used with children, and if so, whether learner types differ, and whether children show correlations with vocabulary, as in studies of spoken language. Positive results would indicate that (i) the link between vocabulary size and processing speed is a modality general aspect of language behaviour in children, and (ii) that the methodology could be valuable for future studies of sign language acquisition. Beyond these core points, however, I don't see the rationale for a number of other things that have been brought into the paper.*

We appreciate the point about how to best frame our findings. We followed your later suggestion and restructured the sequence of the results (thank you). We have expanded the introduction with the goal of focusing the paper on a more streamlined set of issues. Please see below for a point-by-point response to your thoughtful comments.

> *As one example, I don't see the argument that immediate gaze shifts constitute an important skill for children to learn (see p. 5) and I'm not sure the paper actually addresses issues from this angle anyway. For example, if children waited an additional 100 ms or more until the end of a sign before shifting attention to visually-present object, would this be expected to change the course of language acquisition in some meaningful way, or have important implications for nature of lexical or referential processing?*

Thank you for pointing this out. We have removed the discussion about any potential implications of rapid allocation of visual attention for learning since this was not the focus of the current study.

> *Similarly, if children's gaze shifts only occurred after a sign was complete, would researchers really take this as evidence against incremental interpretation? I believe an equally likely conclusion would be that children's ability to redeploy gaze from an attentionally-attractive location in visual space (a moving person) to another location is simply slower than in adults, making the gaze tracking methodology ill-suited for this age group. So the rapid gaze-shifting ability strikes me as really being a methodological issue in relation to what the authors want from the eye movement measures, not a theoretical one.*

This is also an important point. We agree that waiting until the end of the sign or the end of the sentence would not provide evidence against incremental processing since there could be other causal explanations for that pattern of data. With that said, we do think that our results -- that shifts tend to occur prior to sign offset -- do provide positive evidence in favor of an incremental account. We think that it is important to make this point since there is only handful of studies showing evidence for incremental ASL processing, and this is the first to show it in young ASL learners. Therefore, we chose to keep some of the discussion about the theoretical implications of our results for incremental ASL processing in the paper. Following your later suggestion, we do try to emphasize the importance of this finding for validating the inference that our task is measuring speed of lexical access, thus setting the stage for our later group comparisons and individual difference analyses.

> *Another statement that (respectfully) doesn't really seem to fit with the study at hand is on p. 18: "Parallel looking patterns for deaf and hearing ASL learners suggest that both groups are sensitive to modality-specific constraints of processing a sign language". Is this really what the results tell us, and what constraints are at issue here? Linguistic reference to things in the here-and-now (which sets up the relevant visual competition in the current case) instead of "absent" referents doesn't strike me as a defining feature of sign language use. Further, even in face-to-face spoken communication, past research tells us that listeners like to look at talkers' faces (although this is clearly not how spoken language eye tracking studies are typically conducted). I'm afraid I don't see the modality-specific argument.*

Thank you for highlighting that we could have been clearer about what we mean by "modality-specific" in the paper. Our point about modality-specific constraints is that during sign language comprehension the speaker is the only fixation location in the visual scene that provides information with respect to the listeners goal of understanding language. In contrast, people listening to spoken language could choose to look elsewhere and still be processing language via the auditory channel -- that is, they can look while they listen.

Thus, we wondered if hearing children's access to auditory information in the daily lives might change how they choose to allocate attention during ASL comprehension -- perhaps they would distribute more looks to the objects over the course of the trial and less attention to the signer driven by their experience using hearing to monitor the environment in their daily lives. On the other hand, deaf children have to constantly use vision to monitor their environments and constantly switch attention between signers and the nonlinguistic visual world (e.g., Lieberman et al., 2014). So it is possible that deaf signers would show different looking behavior. We present one plausible alternative hypothesis -- that they might wait until the very end of the sign.

The parallel results suggest that learning ASL as your first language leads to similar patterns of behavior in terms of the time course of allocating attention between the signer and the objects in the world. This provides evidence against an account that access to auditory information plays a role in this particular behavior.

> *One suggestion would be to stick to a slightly more streamlined set of issues, and to reorder things so that the question of whether children can shift gaze before a sign is completed sits in the #1 spot (instead of #3). Once this point is firmly established, I think the remaining points and their corresponding measures will have a more logical place in the flow of the argument, e.g., now that we know the gaze shifts are occurring as signs unfold in time, it makes sense to explore differences in this real time processing across participant groups and potential correlations with vocabulary size. Indeed, on pp. 19-20 the authors themselves make the point that their finding that participants can quickly shift attention away from the speaker validates the inference that timing measures can reflect the speed of lexical access. So I do think there would be value in reordering the way in which things are presented to make the value and role of the different measures more apparent.*

We sincerely appreciate this suggestion, and we have restructured our results section to reflect the order that you suggested. We think that this new sequence is much clearer. We also tried to focus our interpretation of the sign offset analyses on both validating the method in order to better interpret our subsequent analyses and on the theoretical claim about incremental ASL processing. We do think that our results provide positive evidence of incremental ASL processing (even if we would not interpret shifts after noun offset as evidence against incremental processing), so we chose to keep this discussion in the paper.

> *2. The introductory section might benefit from some more consideration of when eye movements reflect lexical vs. simply referential processing. This feels a bit fuzzy at times. The challenge, of course, is that eye movement paradigms make use of an overt referential behavior (linking an auditory or visual symbol to a real-world referent) as a measure for making inferences about lexical processing. But I think a bit of rewriting can sharpen things up.*

Thank you for this suggestion. We added more discussion of the linking hypothesis to the introduction. We try to make it clear that one of the goals of this work is to provide evidence that ASL users' eye movements to named referents reflect speed of lexical processing and not some other process (p. 3-4).

> *3. The motivation for the comparison between hearing and deaf native signers could be clearer. I think there is definitely value to this comparison, but the motivation as currently stated comes across a bit weak. The authors speculate that there may be differences based on the groups' differential access to auditory information. But why*

> *exactly would this matter when performing what is essentially a purely visual task? Is the specific idea that the hearing children could be more susceptible to auditory attentional capture (assuming the testing environment, like their home environment, is not totally free from environmental noise)? Or could differences arise because hearing children might be "splitting their time" across ASL and English and as such be less fluent compared to monolingual ASL learners even when matched on certain measures?*

The analysis is motivated by previous research that uses hearing native signers as a way to separate effects of auditory deprivation from those of learning a visual-manual language. For example, Bavelier, Dye, and Hauser (2006) review evidence that deaf individuals show enhanced sensitivity to motion in the periphery, but hearing signers do not, suggesting that this effect is driven by a lack of access to auditory information. In contrast, other research has shown that both hearing and deaf signers exhibit increased mental rotation abilities compared to hearing spoken language users, suggesting that learning ASL is driving this effect.

In the current work, we present two hypotheses: (1) that deaf children's daily experience relying on vision to monitor the linguistic and nonlinguistic visual world might change the timing of eye movements in response to language in our task. Specifically, we thought that young deaf signers might be slower to disengage from the center signer in order to reduce the chance of leaving early and missing the upcoming linguistic signal. Or (2) it is possible that the experience of learning a visual language and the coupled with the "in-the-moment" constraints of processing a visual language would lead to similar looking patterns regardless of hearing status.

The parallel results for deaf and hearing signers provide preliminary evidence in favor of the second hypothesis, suggesting that acquiring ASL as a first language leads to similar dynamics of eye movements during ASL sentence comprehension regardless of access to auditory information in children's daily lives.

> *Also on this point, I think speculations for possible differences for the two younger groups may need to be discussed separately when it comes to the question of whether gaze can be rapidly disengaged from the speaker vs. the question of whether signs are processed rapidly/incrementally. Measurement issues aside, one could imagine, for instance, that native deaf signers might be worse at the first but better at the second.*

Thank you for the interesting suggestion. We have added these different explanations as a paragraph in the discussion.

> *[Related to this: p. 23:19 "...while the hearing children could use vision and hearing to process incoming information." This gives the impression that there was speech information in the current stimuli. There wasn't, was there?)*

Thank you for pointing out that we could have been more clear in our language here. There was no auditory information in the stimuli. We intended to refer to hearing children's access to auditory information in their daily lives. We clarified our language in the text.

> *4. There were a few methodological details that were not clear to me. Did each participant encounter 8 trials in total?*

Participants completed 32 test trials, four trials for each of the eight target signs. We also included 5 filler trials (e.g. "YOU LIKE PICTURES? MORE WANT?") interspersed in order to maintain children's attention (see Procedure on p. 10).

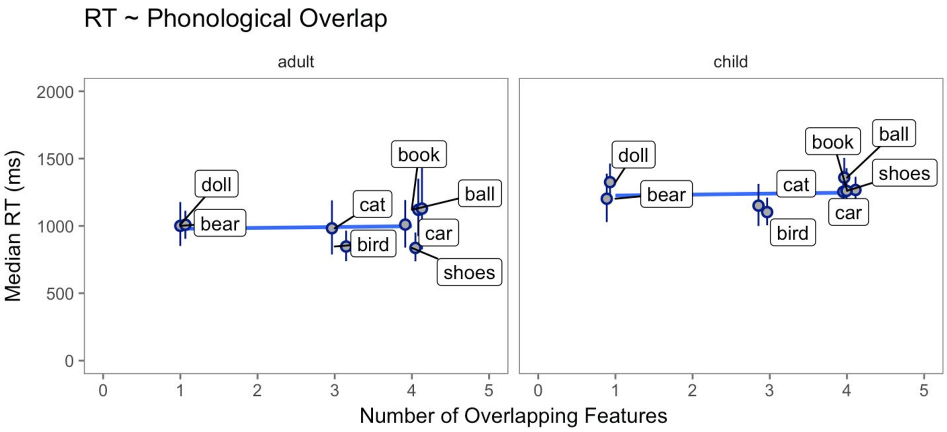
> *Were they balanced for the two different talkers and the two different question types, or were these cycled across different subsets of participants? (Did every participant see exactly the same stimuli?)*

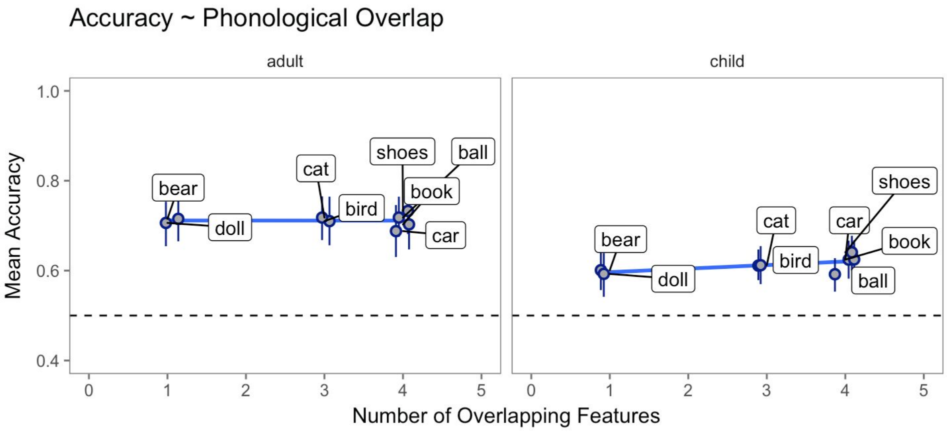
Thank you for pointing this out. No, every participant did not see the same exactly the same stimuli. Each participant saw one stimulus set with a single signer and question type. Participants were roughly evenly distributed across the two different signers and question types. 16 children saw the sentence-initial wh-phrase structure and 13 children saw the sentence-final wh-phrase structure. We did not find evidence of differences across the two stimulus sets. We have made these points clearer in the paper.

> *Also, although the authors note that the item pairs have "minimal phonological overlap" (p. 10), I think readers will want to know about the precise extent to which the various target and competitor signs were distinct with respect to ASL parameters like handshape, orientation, location, etc.*

Based on this suggestion, we used ratings of phonological overlap taken from a recently created database of lexical and phonological properties of nearly 1,000 signs of American Sign Language (Caselli et al., 2017) to quantify this feature of our stimulus set. Our target-distracter signs vary in the degree of phonological overlap from 1-4 parameters. We have updated the methods section to include this information.

Here is a plot showing RT and Accuracy as a function of the number of overlapping features in the item pairs.





We did not see evidence (in our dataset) that degree of phonological overlap affected the timing or accuracy of looking behavior. One possibility is that our empirical definition of sign onset reduced the chance that phonological overlap was a factor in affecting the time course of sign comprehension.

We included this analysis in the online supplementary materials.

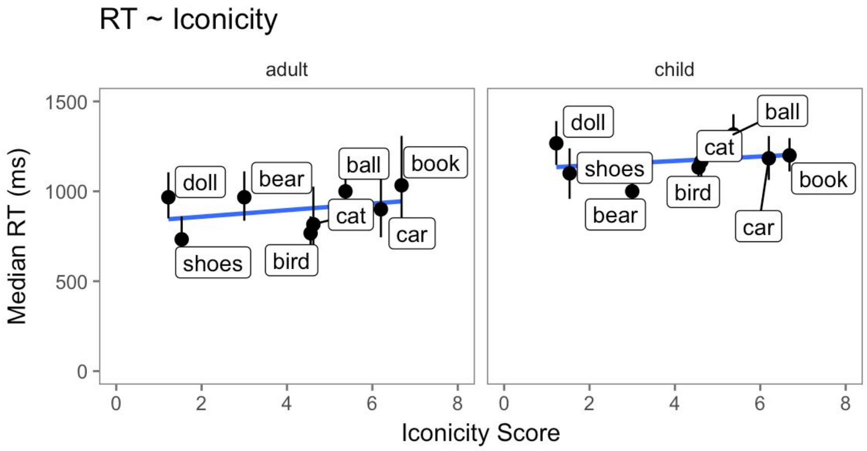
> *There were also some other places where there could be more clarity in describing the stimuli (e.g., p. 10 line 55: the earlier examples made it sound like the target sentences are in fact questions. Was there an additional question after those questions?)*

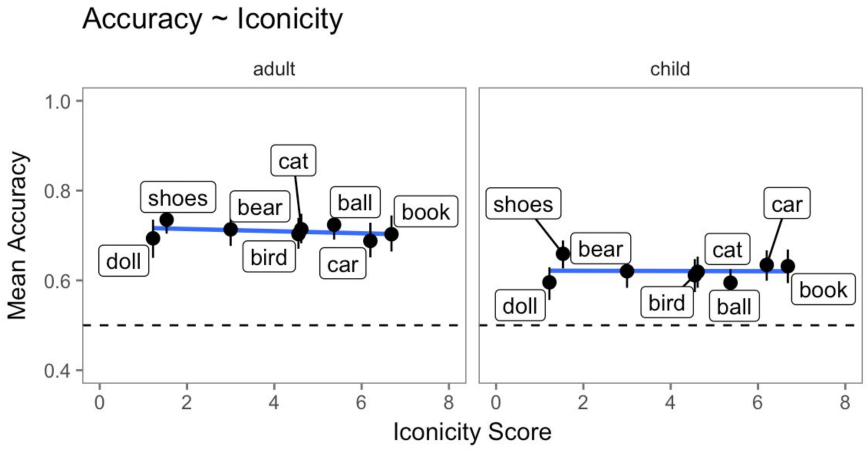
We have added more information throughout the methods section for clarification.

> *5. One methodologically and theoretically important issue I would like to see addressed concerns the issue of iconicity in the signs used as stimuli. The sign for ball, for example, seems to have a clearly iconic aspect to it, reflecting shape characteristics that can be easily mapped to the target object on the basis of low-level attentional features. Was this true of other signs as well? This seems important*

because iconic signs could plausibly exaggerate the speed and/or accuracy of processing. If measures of iconicity are available for ASL signs, one possibility would be to include iconicity as an item-wise predictor in the statistical analyses.

Thank you for pointing out that iconicity could be important for accurately estimating the lexical processing in ASL. To address this question, we measured the association between iconicity ratings and both Reaction Time and Accuracy. The iconicity ratings were taken from a recently created database of lexical and phonological properties of nearly 1,000 signs of American Sign Language (Caselli et al., 2017). We did not see evidence of an effect of iconicity on either processing measure. Here is a figure showing RT and Accuracy as function of iconicity of the target signs.





We included this analysis in the online supplementary materials.

6. Re: measures of interest. The operationalization of "sign onset" seems quite different from "word onset" in studies of spoken word recognition (and is perhaps closer to 'uniqueness point' than word onset?). The initial speech sound in a word is rarely, for example, sufficient to entail 100% agreement across listeners regarding target identity, in part because of the nonlinear nature of speech cues (e.g., with information about the place-of-articulation of consonants being carried by vowel *formant transitions, etc.). I understand the challenges here ,but it seems relevant to consider whether the sign onset measure as implemented might exaggerate measures of incremental interpretation (proportion of target processed), or alternatively whether "incremental" (rather than "rapid", etc.) is really the right term to describe the processing of information that is already fully available at the left margin of the measurement interval. Also: how exactly was this sign onset measure calculated? Were the adult signers watching videos in slow motion, or via a gating task?*

Thank you pointing out that we could have described the gating task more clearly. We added more details about the task in the methods section.

p 14, sentence beginning "In studies with adults...". This sentence might need a tweak or two, as the truth of this statement really depend on the design. If for example there is a phonological competitor in the display (e.g., candy + candle), then initial fixations can simply reflect early lexical hypotheses, not the speed of lexical access for the intended target.

This is a good point. We have added a footnote to clarify that we are only referring to visual world paradigm studies without phonological competitors.

It would be good to know the rationale for using an analysis window for the accuracy measure that ranges from 600-2500 ms after the onset point.

We agree that the choice of analysis window is important in this study. We selected 600-2500 ms based on the middle 90% of the empirical RT distribution. We added this justification to the description of our accuracy measure (p. 13). We also worried about whether the results were sensitive to this analysis decision, so in the online supplement we include a sensitivity analysis where we vary the upper bound (+/- 300 ms) of the analysis window and show that our findings are robust to this choice.

The authors report separate statistical models using age vs. vocabulary predictors is because this predictors are highly correlated. I think it likely makes sense to report this correlation somewhere (maybe I missed it).

Yes, we report separate models because age and vocabulary are highly correlated. We have added the correlation to the paragraph addressing this limitation in the General Discussion (p. 24).

7. Re: Discussion section.
i. Bottom of p. 25 and highlights: "...as soon as listeners have enough information ...". Is this perhaps a bit too strong, given that "sign onset" was identified as the point in

the video at where there was sufficient information for target identification yet in absolute terms gaze shifts tended to occur toward sign offset?

Thank you. We have changed this phrasing throughout. It now reads "... as signs unfold in time and prior to sign offset." We think this stays closer to the measured behavior.

ii. How do the results for adults compare to the Lieberman et al. study cited in the introduction? (p. 4:36)

Thank you for suggesting this idea. The closest comparison in Lieberman et al. (2013) is their Unrelated condition, which consisted of a target picture and three competitor pictures whose corresponding ASL signs shared no semantic or phonological properties with the target sign. Adults' average latency to shift gaze away from a center fixation to a named object was 844 ms. This is strikingly close to our measure of adults reaction time: 862 ms. We added this interesting comparison to the Discussion (p. 25).

iii. I would urge the authors to be just a bit more cautious when describing the age-related differences in reaction time in the discussion. This is because there was no control condition that could assess the possibility of age-related gains in the efficiency of controlling visual attention to scene regions.

This is a good point. We have added it as a paragraph in the limitations section of the discussion.

8. MINOR POINTS
p. 3:27. Introduction: The description of adult patterns in studies of spoken word recognition (i.e., "as soon as the auditory information is sufficient") sounds as if it applies to infants, whereas in reality infants are measurably slower and show delays in target identification in relation to the uniqueness point within a word (as the authors note elsewhere).

We changed the phrase "as soon as" to "soon after." We think this stays closer to the developmental findings in spoken language.

p. 4:19. (Sentence beginning "As in spoken language..."). Most current models of spoken word recognition avoid talk of 'stages' and instead favor the notion of cascaded processing. It may be helpful for the authors to clarify what they mean here (or just eliminate this wording if not needed to make a specific point).

Thank you. We removed the word "stages" from this sentence.

p. 4:57: "...less of the linguistic signal" Please clarify for the readers (e.g., does this

> *mean a shorter time sample of the unfolding linguistic signal?) Also, the description of the different gating tasks could be clearer for people unfamiliar with this methodology. As it stands, "increasingly longer segments" risks sounding like the participant hears a word or sees a sign in its entirety several times, but each time more slowly / stretched out.*

Thank you. We have clarified the description of the gating task and the meaning of the phrase “less of the linguistic signal.”

> *p. 5:52: The allocation of visual attention is not a 'basic learning mechanism'. Please clarify what is intended here.*

Thank you for pointing this out. We changed the the term “mechanism” to “processes” and also now use the phrase “coordination of joint visual attention” which is the construct that we had originally intended to discuss in this paragraph, and stays closer to the behavior measured in Lieberman et al. (2014).

> *References section: A Lieberman et al. study is listed twice in the references, with two different dates. A (different?) Lieberman et al. study mentioned in the introduction seems to be missing. Also, some references do not conform to APA style (e.g., handbook entries, capitalization rules), and some references run together in a few places.*

Thank you for pointing this out. We have fixed both Lieberman references, and we updated the references to conform to APA style.