Dear Dr. Hulme,

Thank you for the thoughtful comments on our manuscript, "Real-time lexical comprehension in young children learning American Sign Language." Please accept our resubmission. We have addressed your comments and the comments of the reviewers, and we believe that the manuscript is substantially improved. Please find below a point-by-point response.

A question of concern to all the reviewers was the theoretical framing and impact of this work. So we would like to highlight the addition of a new, exploratory analysis of the timing of children's and adults' gaze shifts relative to the offset of the target signs. This analysis shows that, despite the presence of competition for visual attention, young ASL-learners and adult signers generate saccades before the end of the target sign and well before the end of the utterance. This finding, along with the links to vocabulary development, provides evidence that rapid, incremental processing of linguistic information is a language-general phenomenon that supports language acquisition regardless of modality.

We'd also like to highlight the addition of several important pieces of information that we hope address concerns about our statistical approach.

1. We provide a more thorough explanation of the link between the data generating process (gaze patterns over a fixed time window) and the choice of statistical model (linear regression, which assumes continuous, normally-distributed outcome variables). Specifically, we show that our mean accuracy and RT measures are normally distributed at the participant level (see figure 1 later in this document), making us more comfortable with using the linear model to estimate developmental change.
2. We follow the recommendation from Reviewer 3 to use a logistic transformation on our accuracy data (mean proportions), making a linear regression more appropriate for these data.
3. We added a parallel set of non-Bayesian analyses to the supplemental materials since to show that these results are robust to choice of analysis framework and since readers might be more familiar with interpreting these results.
4. We removed the median split analyses, removing unnecessary hypothesis testing that distracted from our original intent, which was to emphasize an estimation approach. We think this version shifts the focus to quantifying the uncertainty in our estimates via interpretation of the full posterior distribution over the coefficients in our models. We think that this approach fits well with Psychological Science's emphasis on reporting interval estimates and will help move the field away from focusing on binary reject/accept outcomes more broadly.

Finally, we would like to point out that this is the first study to use high-resolution, eye-tracking measures to assess language development in young visual language learners. It is also the first study to characterize the looking behavior of both native deaf and hearing children as the process ASL in real-time. And while there have been diary studies of language production in these populations, the development of these precise quantitative measures lays the foundation for future work exploring parallels and differences between the psycholinguistics of visual and spoken languages.

Please do not hesitate to contact us if you have any questions or concerns. We look forward to your consideration of this revision.

Sincerely,

Kyle MacDonald

Editorial

> *The first concerns the theoretical framing and interpretation of your study. The reviewers make a number of clear and incisive comments on this (and two of the reviewers question whether the theoretical impact of the work is sufficient to warrant publication in PSCI). It is clearly important that any revision successfully addresses these concerns.*

We sincerely appreciate this feedback, and we now highlight why studying eye movements during sign language processing provides an important theoretical advance for the field. Specifically, we emphasize what we think is a reasonable and interesting alternative hypothesis: that the competition for visual attention present in ASL processing might change how incoming linguistic signals drive gaze shifts, thus complicating the link between this behavior and the underlying construct of interest -- speed of lexical access. We then provide evidence against this hypothesis with a new, exploratory analysis showing that children and adults generate saccades away from the center signer prior to the offset of the target sign. Together, we think that the new framing and analysis strengthen the manuscript and will increase the theoretical impact of the work.

> *The other concern that stands out relates to your analyses. Again all of the reviewers address this issue though their concerns are quite varied. The most radical suggestion (from Reviewer 4) is that you drop your Bayesian analyses all together. It is up to you to judge how best to deal with these comments on your analytic approach, but it is clear that major changes to the way you have analysed your data are called for.*

We appreciate these critiques about our analysis decisions, and we have made several changes to the description of our analysis (see "analysis plan" section). In response to several Reviewers' questions about the use of linear models given the data structure, we added a paragraph justifying this decision. We also provide the distribution of the two VLP measures in the cover letter below. Moreover, we followed Reviewer 4's suggestion to remove unnecessary hypothesis testing and focused our results on visualization and presenting interval estimates of our effects. Finally, in response to Reviewer 3's discomfort with the Bayesian approach, we added a parallel set of non-Bayesian analyses to the supplemental materials to provide readers with a more complete set of information to help interpret this work.

Reviewer 1

> *The important contribution of the study is to show that the processing factors that underlie language acquisition are robust to what, on the surface, appear to be roadblocks to accurate and speedy word recognition. However, this is not how the present study is framed in either the introduction or discussion.*

We sincerely thank you for this suggestion. We restructured the introduction and discussion to place more emphasis on why extending the work on eye movements during language comprehension to a visual-manual language like ASL is of theoretical interest. We highlight the point that, despite high competition for visual attention, incoming language still drives rapid gaze shifts to named objects. We also included a new, exploratory analysis that attempts to directly quantify the amount of linguistic information (i.e., proportion of sign presented) that children and adults processed before generating a saccade, providing evidence that ASL users shift prior to the offset the target sign.

> *The introduction emphasizes the possible differences in the sublexical organization of signs in comparison to spoken words as motivating the study, specifically the notion that the sublexical features appear simultaneously in the visual sign signal but sequentially in the auditory spoken word signal. The one paper cited showing this to be the case for ASL is dated, and tested few subjects with few lexical items without the benefit of modern technology. Subsequent work has demonstrated a temporal unfolding of sublexical features in signs, but this more recent work is not cited or dealt with. Curiously, the present paper abandons this hypothesis throughout the methods and results and instead describes sign recognition in the child subjects as one of temporal unfolding. (The cited studies on sign language processing and the development of lexical processing in hearing children do not reflect the status of research in these fields.)*

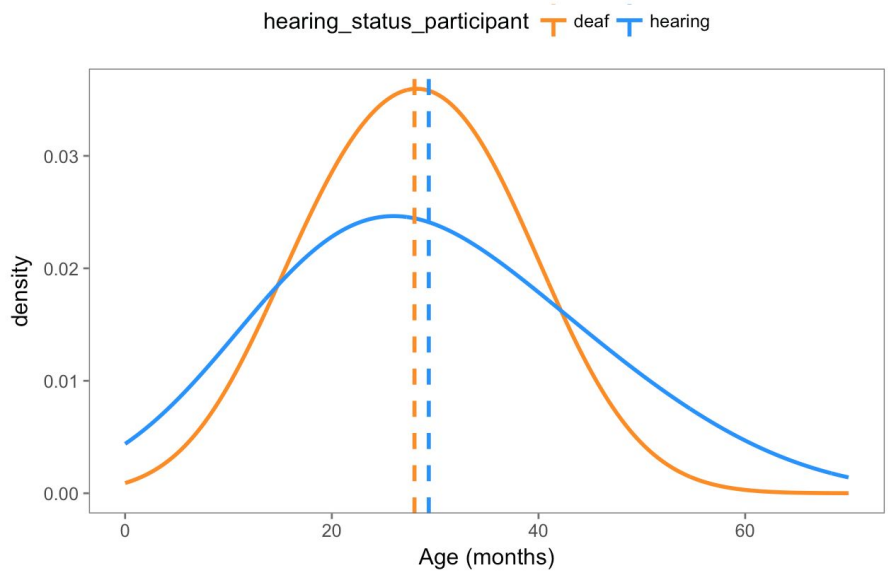
Thank you for pointing us towards this literature. In the introduction, we added citations to recent work on lexical access in sign language that addresses the debate between simultaneous vs. temporal unfolding, including: Gating work from Morford and Carlsen (2011); EEG work from Gutierrez et al. (2012); Computational modeling work from Caselli and Cohen-Goldberg (2014) and Chen and Mirman (2012).

> *All the explanation about cochlear implants and a dwindling sign language population in the present ms are both a distraction and a speculation. I am unaware of evidence that deaf parents have ceased using their own language (ASL) with their deaf and hearing children. (Note also that the cochlear implant issue does not apply to hearing children nor change the proportion of deaf children with deaf parents; moreover since the authors found no differences between the deaf and hearing groups, deaf children who have CIs whose parents use ASL with them could have been included in the study).*

We appreciate this point, and we streamlined the justification of our sample size by removing the speculation about the effect of Cochlear Implants.

> *Regarding the reporting and the analyses, because age is a key factor in the effect, it is important to provide a table giving the ages of deaf and hearing children separately.*

Thank you for this suggestion. We agree that it is important to provide more descriptive information about the sample, so we added a table (Table 1) with the age distributions for both the deaf and hearing children to the results section. Moreover, we also provide a plot of the age distributions for the hearing and deaf ASL-learners, showing the high overlap between the two. We would be happy to include this plot in the manuscript if the editor or reviewers felt it would be helpful for the reader.



> *An equal distribution of hearing children across the age bins would show that the results are not being driven by the hearing children. They could have acquired larger vocabularies by virtue of overhearing spoken English (some of the children were as old as 4 years of age), and the present results find vocabulary to correlate with processing speed. This possibility is not mentioned in the present ms. No reason is given as to why the English vocabulary of the hearing children was not measured. The claim is that they are monolingual, but evidence for this claim is not given.*

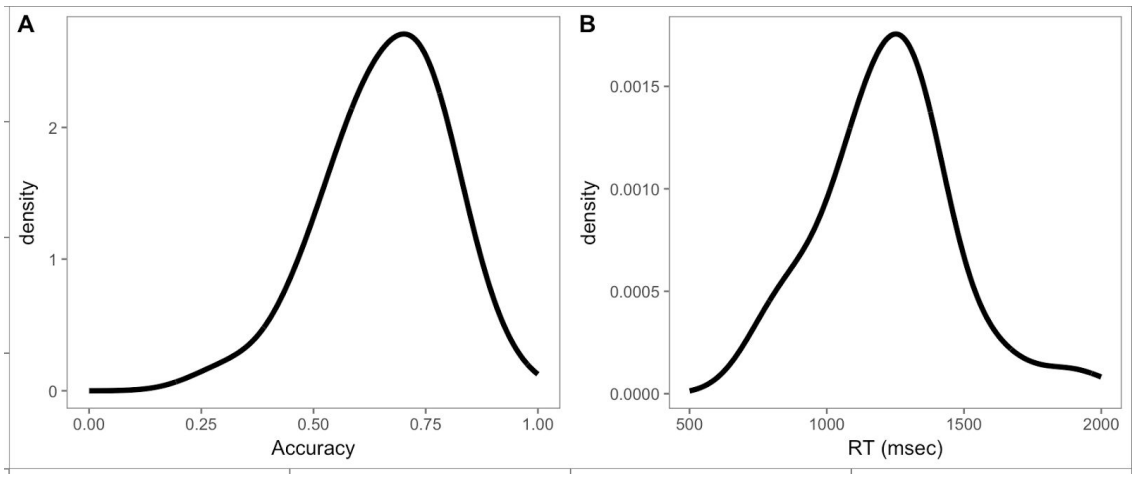
This is an important point for our argument. We added information about the distribution of hearing children across ages to the new demographics table and to the scatter plots showing relations between processing and age/vocabulary. We hope that this will make it clear that the age distributions of hearing and deaf children are quite similar. Finally, we reiterate that all children, regardless of hearing status, were exposed to ASL from birth and experience ASL as their primary language in the home and at school. In conjunction with the results of our first analysis showing evidence for no group differences in the speed and accuracy of looking behavior between the young deaf and hearing signers, these factors provide additional justification for treating the hearing children as monolingual ASL-learners and grouping data across the hearing and deaf ASL learners.

> *Regarding the reliability measures, they are reported as proportion of agreement for video frames, but whether these were isolated video frames or over an entire trial or subject is not described.*

We updated the description of the reliability measures to make this clearer. It now reads, “Agreement was scored at the level of individual frames of video and averaged 98% on these reliability assessments.”

> *Regarding the accuracy analyses, two alternative pictures were given for each trial with a 50/50 chance of being correct. Here a Gaussian distribution is used, but accuracy would be a value between 0 and 1. The number of items (n) should be included with the use of a binomial distribution (Bernoulli (p), where p is determined by the intercept and the slope of the factor of interest) for each item to obtain an observation for that item (either correct or incorrect) before obtaining mean accuracy. It is not clear if the number of test items was included in the model.*

There is a lot of discussion in the field about the best way to analyze the accuracy of eye movements in the visual world paradigm (e.g., Barr, 2008; Jaegar, 2008; Mirman, 2014) in order to deal with issues related to the non-independence of time points and the underlying structure of the data. Here, we took an approach that is standard in our sub-field where we average binary outcomes at the 33ms level over an analysis window across multiple trials to get a mean proportion looking to the target score for each participant, making the measurements independent of one another. These scores represent the probability of looking to the target picture across all trials for a given individual and are relatively normally distributed (see Panel A of the plot below). We also follow suggestions from the Reviewers to use a logistic transformation on the accuracy variable (proportion looking) to further justify using the linear model to estimate the relations between age/vocabulary and accuracy/RT.



> *Regarding the RT analyses, all the figures show the window of analysis, but not the window of analysis in relation to the onset and offset of the stimulus sign and preceding signs. Without this information, it is impossible to determine when children are looking away from the sign stimulus toward the pictures. An important piece of information for the field is what proportion of the sign signal the children needed to see before looking away.*

We really appreciate this suggestion and were inspired to include a new analysis of children’s and adult’s first shifts relative to the offset of the target signs at the trial-level. See the “Evidence for incremental processing of ASL” sub-section in the Results section. This analysis provides evidence for incremental processing of ASL in young children -- that is, although they require more of the sign compared to adults, they do not wait for the end of the target sign (or the entire utterance) before

generating a shift away from the language source. We agree that this is important information for the field because it indicates that the timing of these gaze shifts is linked to the underlying process of lexical access and thus can be leveraged to measure individual variation in processing skill. Moreover, these findings are of theoretical interest since they indicate that the rapid influence of language on visual attention is a language-general phenomenon.

We also added information about median sign offset to Figure 2, showing that proportion looking to target curves start to increase well before the offset of the target sign.

> *Regarding the vocabulary measure, what was the name of the vocabulary inventory used; why was expressive vocabulary rather than comprehension vocabulary used. Presumably lexical processing speed is more correlated with lexical comprehension than production in young children.*

We developed the vocabulary inventory specifically for this project. We structured the instrument based on the Macarthur Bates Communicative Development Inventory (Anderson & Reilly, 2002), however, we needed to adapt the instrument to be appropriate for the wide age range of participants tested. We focused on production, rather than comprehension, because previous studies have shown that caregivers are not able to report on comprehension in children older than about 1 ½ years old (Fenson et al., 2007). Finally, several studies using the looking-while-listening task with spoken language users have shown strong correlations between efficiency of real-time language processing and vocabulary production (e.g., Fernald et al., 2006; Fernald & Marchman, 2012).

Reviewer 2

> *Concerns regarding the paper include presentational issues, the fact that for some of the analyses age was (apparently) dichotomized,*

We appreciate this point since median splits create arbitrary age categories and reduce statistical power. We removed the dichotomization of the children age from our visualizations and now just plot children's performance against adults' performance. Moreover, we make it clearer that we treat age as a continuous variable in the linear models estimating developmental change.

> *The project uses an individual differences approach, looking at the relationship between age and the various eye-tracking measures. Yet no information about the internal reliability of these eye-tracking measures is provided. Standard measures would include coefficient alpha (~the mean of all split half reliabilities).*

Thank you for this suggestion. We added internal reliability measures (Cronbach's alpha) to the results section.

> *A picture with the example display is needed. It is hard to think about the task without seeing an example display with the size and location of the critical pictures as they were in the task.*

Thank you for pointing this out. We added an image of the stimuli layout to Figure 1.

> *The gray on gray in Figure 2 is really difficult to parse. It would help a lot if the authors got rid of the gray shaded area in the background and indicated that information elsewhere, perhaps on the x-axis. (same comment goes for Figure 3).*

Thank you for pointing this out. We reduced the size of the grey shaded area to only appear below the proportion looking curves. We think this shows the analysis window without obscuring the curves.

> *The HDI is going to be unfamiliar for most readers; guidance in interpretation of these values is needed. i.e., explain what it means for the HDI to include zero. Same goes for the BF.*

We added information about how to interpret these values in several places, including the analysis plan and the caption of Table 2.

> *This is the first time the authors mention the fact that the children have been put into different age groups. Continuous variables are best treated as such; dichotomizing a continuous variable leads to a loss of power and can generate spurious findings (see Preacher, et al. 2005, Psychological Methods, for multiple citations and some discussion of why dichotomization is a problem). Further, there is no justification for why 27m is used as the cutoff.*

We agree and thank you for pointing us to this literature. We removed the median split analyses and we treat age as a continuous variable in the linear models estimating relations between age/vocabulary and processing efficiency.

> *The accuracy measure is a proportion based measure over a 1900ms time window. It is likely that these distributions were highly non-normal; some discussion of the appropriateness of the analysis technique is needed given the data distributions.*

As we indicated above in our response to Reviewer 1, our approach is standard in our sub-field where we average binary outcomes at the 33ms level over an analysis window to get a mean proportion looking to the target score for each participant. These scores are normally distributed (see Panel A of the plot above), and hence, are appropriate for a linear modeling approach.

> *Figure 4, left panel is described as plotting "Accuracy" but the y-axis is plotting proportion looks to the target and distractor. If this is a measure of accuracy it should only plot target looks, based on the definition of accuracy on p11.*

Thank you for pointing this out. We removed the proportion looks to the distractor from the figure since this measure is not a focus of the study, making the y-axis is more informative.

Reviewer 3

> *In typical eye-tracking studies, the eye-movements can simply be treated as largely a read out of the activation state of the lexical items. However, here they serve two roles -- first they are the only route by which the input (the word) can be obtained; second, they are a readout of activation states. Consequently differences in fixations could derives from how long the participants need or want to look at the ASL stimulus or from how rapidly they activate the correct word.*

We agree with this point. However, we think that the competition for visual attention between the linguistic signal and the nonlinguistic visual world is one of the reasons why understanding eye movements during real-time ASL processing is such an interesting case study. That is, it could have been the case that the incoming linguistic signal would not drive rapid shifts in visual attention prior to the offset of the target signs, thus complicating the link between eye movements and the underlying construct of interest: speed of lexical access. But the current work provides evidence that both adults and children on average shift prior to the offset of the target signs, suggesting that these eye movements do provide an index of speed of lexical access. More work is needed to understand the unique contributions of speed of lexical access and the other factors the reviewer mentions, and current studies in our lab are trying to tease these apart.

> *Very little information is reported about the deaf adults who were tested. In particular, it would be important to know the age of onset of deafness, the way they learned ASL (did they have deaf parents, at a deaf school), when ASL exposure began, and if they were profoundly deaf or had any residual hearing (and whether that was tested), or were actually hearing (as many of the children were).*

Thank you for pointing this out. We added information about the adult sample, which only included deaf adults who had lost their hearing early in life and were exposed to ASL from their parents or from school. We did not collect information about residual hearing.

> *As reported, the statistical model includes hearing status (which I think is only relevant for the children, but I can't be sure since it was not reported for the adults), age, and vocabulary. This was stated at the outset of the intro, but it may not be entirely true for all of the submodels (though it was very hard to tell) as the individual models were not described.*

Yes, the hearing status variable is only included in the models estimating developmental change in speed and accuracy within the children sample since all of the adults were deaf. We also added information about hearing status to the scatter plots showing relations between processing efficiency and age/vocabulary, so the reader can see that the hearing and deaf children were evenly distributed across the range of ages in the sample. Finally, the adult data were not included in the linear models and serves only as an estimate of the developmental endpoint to which children are making progress towards.

> *First the typical rule of thumb for between subject regression designs is 10-15 participants for factor (this is meant to avoid inflating R2 values, and increasing Type I errors). With*

only 45 participants (and I understand why the sample needs to be that small), or 29 deaf children, this is really pushing the lower bound of a three parameter model. While I realize the authors are using Bayesian models, as I understand it, the problem really comes from estimating the linear model, not the evaluation of significance so it likely holds. The Bayesian model would seem to have more free parameters (although it constraints them) so I really am not sure how it all adds up. But this feels like a lot of model for not a lot of data.

We are not aware of this rule of thumb for estimating a linear model, but we do share your concern about the sample size. The simplicity in terms of number of parameters is one of the reasons we chose the linear model for our analysis. Also, the Bayesian linear model has the same number of parameters as the Maximum Likelihood version (which we also include in the supplement in case readers are more familiar/comfortable with interpreting these analyses). Finally, in our discussion we try our best to emphasize that this study is the first data point in what we hope will be a growing area of work exploring the developing psycholinguistics of ASL using online processing measures, thus emphasizing the cumulative science necessary to have confidence in any finding.

Second, it is unclear if any of the adults are hearing or not (like the children). If they aren't hearing status becomes quite confounded with age (though it is possible that with proper coding of the contrasts this wouldn't be a problem, and some of the models don't seem to include the adults).

All of the adults in our sample were deaf, and they were not included in the linear models estimating developmental change.

Third and similarly, vocabulary is almost certainly confounded by age, and it was not even clear how or if this was assessed with adults. The write up seem to treat these separately, but it was not clear if both were in the model or not. If they're separate then their independent effects cannot be assessed (as the authors acknowledge in the limitations); if they're together, its not clear what to make of it. Here a residualization strategy could really help.

Thank you for pointing out how we could be clearer in describing our analysis. To accomplish this, we added a new "analysis plan" section where we highlight the logic our analysis, including the link between the structure the data and the type of model used to estimate developmental change. To be clear, adults were not included in the analyses of developmental change in either age or vocabulary. We also emphasize in the limitations section that age and vocabulary are highly co-linear and therefore relations to processing efficiency were analyzed separately. Future work in our lab is planned to explore these relations in children who fall within a narrower age range, allowing us to assess the independent effects of each factor.

Fourth, it was unclear if a linear model was appropriate for the accuracy data (which is proportional). Some type of scaling (empirical logit or log scaling may be needed).

See response to Reviewers 1 and 2 and the figure showing the distribution of mean Accuracy and RT scores above. We also followed your suggestion (thank you) and used a logistic transform on the mean proportion accuracy data to make the linear model more appropriate.

> *Finally, I have to admit some discomfort with Bayesian models in this case. I understand the arguments for them, and I don't disagree. But at the same time, we have a small sample size of highly variable individuals, and a class of models that most people in the developmental community don't understand . It's just hard to know if this data is being held to the same statistically conservative standard as everyone else. This is particularly the case given that the authors appear to be testing the hypothesis that there is a relationship between age/vocab and accuracy/RT and the models are constrained in a way that seems to put a prior on a positive relationship between vocab/age and outcomes. I see the value of Bayesian models in this case, but in the era of false positive psychology, I'd like some reassurance that we're being appropriately conservative.*

Thank you for raising this concern. We took your suggestion and added a parallel set of analyses to the supplement that uses non-Bayesian alternatives. Our hope is that this will provide readers with a diverse set of information with which they can better evaluate the work.

> *The major methodological problem however is that the participant can't be looking two places at once, and here the output measure (the "read out" of lexical activation) is confounded with the input modality (the input to lexical activation). Given this, it is entirely possible that a delayed look to the correct object only seems to be delayed, if it is really a longer look to the ASL input. This could result in a lengthy delay even if lexical activation is really quick. Its impossible to disentangle this which makes changes in RT very hard to understand – do they really reflect differences in lexical processing efficiency? Or do they reflect differences in how long listeners fixate signs? Or are they strategic in some way?*
>
> *Even if these issues could be disentangled, I'm not sure I see the theoretical novelty and impact of this work. The authors raise a few good points as to why lexical development might be different in ASL users. But those would seem to argue mostly that it is delayed, and they mostly speak to vocabulary growth (the difficulty in using social cues or other visual cues to identify new words) and the linking of word forms to semantics; it is not clear if they apply to processing nor to wordform learning (which is often how this paradigm is interpreted. These do not argue for a fundamentally different relationship between processing and outcomes, nor do they argue that processing should not improve with experience / age. So I'm not really sure what the null hypothesis is here and why it is important to rule it out. That's not to say that the work is not valuable--it is. I just don't see it leading to a major insight into language development as a whole or even sign language development.*

We sincerely appreciate your comment about the importance of laying out a clear alternative hypothesis in our paper. In our revision, we emphasized what we think is a reasonable/interesting alternative model -- that competition for the visual channel in ASL could modulate the rapid influence of language on visual attention that has been a hallmark of spoken language processing. That is, it could have been the case that signers would wait until the end of the target signs or even the entire

utterance before generating a saccade to a named referent. This would complicate what we could learn about lexical access in young ASL-learners by measuring this behavior. However, our results provide evidence against this hypothesis and help to lay the foundation for future work using eye movements to understand the psycholinguistics of ASL.

> *Page 4, line 15: "For example, as in spoken language processing, signers are influenced by both lexicality and frequency…" It would be useful to include citations to the relevant studies in spoken language processing.*

We added a citation to Chen and Mirman (2012) who review the relevant psycholinguistics literature from spoken language.

> *Page 5: "Other research has investigated how the visual nature of sign language might influence children's interactions with caregivers and thus affect learning mechanisms…" I don't think I quite understand the point of this paragraph;*

Thank you for pointing out that we could have been clearer with our logic in this section. We think that rapidly establishing reference via looking at objects while hearing labels provides the input (i.e., word-object co-occurrence data) to build a stable lexicon. In ASL, however, this input mechanism is more complex since the eyes have to process both information streams. Our goal with this paragraph (and the following one) was to set up the alternative hypothesis that competition for visual attention complicates the link between eye movements and underlying processes for both language processing and learning.

> *Page 6, line 50: The acronym CODA is introduced here but it took me a bit to realize it referred to "Children of Deaf Adults" (I think because that phrase was embedded in other phrases). It might help to capitalize those letters in the prior phrases.*

Fixed -- thank you.

> *Page 8: "The VLP task yields two measures of processing efficiency, reaction time (RT) and accuracy." This is a minor issue, but accuracy doesn't feel like a measure of processing efficiency; it is much more analogous to a measure of offline knowledge.*

Thank you for raising this issue. We do think that accuracy is also a measure of processing efficiency since it reflects the tendency to orient to and maintain gaze on the named object over a time window, thus providing a continuous measure of graded word knowledge that an offline measure (e.g., pointing or naming) would not be able to provide.

> *Page 9: In the description of the stimuli, it would be quite helpful to report the duration of the ASL stimuli as well as the variability among words – I think this is likely to be quite important for interpreting the real-time processing measures.*

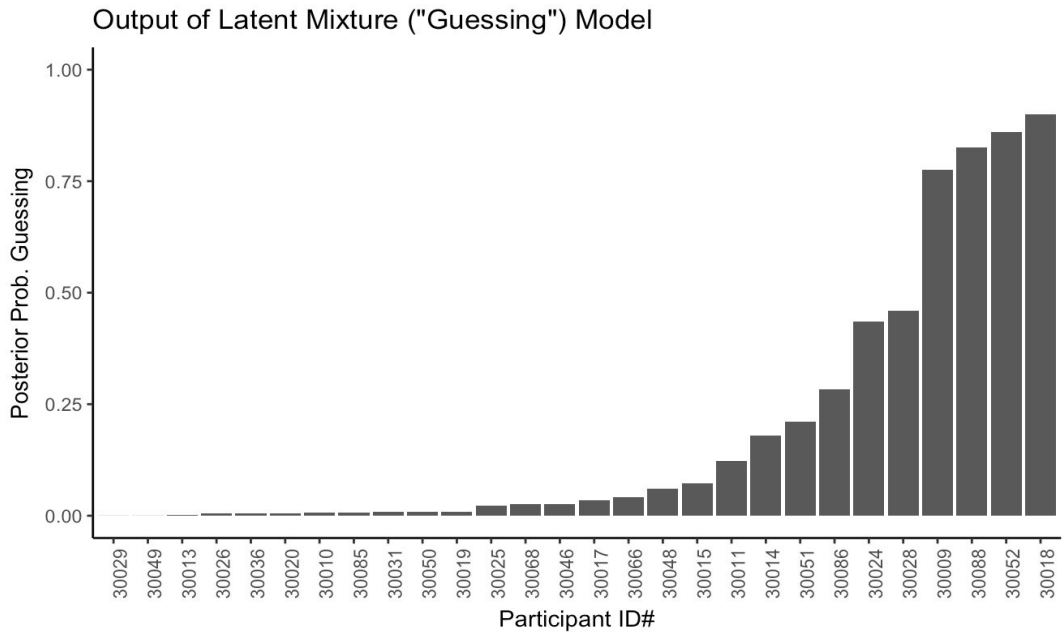
Thank for you pointing this out. We have added information about the median sign duration and the range of sign durations to the methods. We also added the median sign duration to the time course plots to facilitate interpretation of timing of changes in looking patterns.

> *Page 11: "The target sentence was then presented, followed by a question." People who are unfamiliar with ASL, may not realize that there is a distinct sign to indicate a question. If you don't know that this sentence feels quite odd.*

To make this point clearer, we changed the relevant clause to read, "... followed by a common ASL question sign (e.g., WHICH?), ..."

> *It would be quite interesting to know something about the distribution of children as guessers or not (I realize it is probabilistic), as that could be useful information.*

Here is the output of the latent mixture model. For each participant, you see the amount of posterior probability placed on the guessing category, with higher bars indicating a higher probability of producing random first shifts. You can see that the model is confident that some children (e.g., 30018) are guessing and will have less of an influence on the coefficients in the linear models, while others are unlikely to be guessing (e.g., 30029) and will have a larger influence on the linear model estimates.



Reviewer 4

> *Justification 2 (robustness to outliers) is incorrect. Bayesian models are just as sensitive to outliers as other approaches. Justification 1 is irrelevant, as the tested (point null) hypotheses do not seem terribly plausible or interesting. Given that "deaf" versus "hearing" ASL learners cannot be experimentally created and they they will differ in all sorts of relevant ways, the test of a null hypothesis that they are the same is unrealistic. Justification 3 is equally irrelevant because these results are robust to various prior specifications -- in fact, the authors show this in the supplement -- so the prior information is largely irrelevant.*

Thank you for these comments about our analytical approach. We agree that visualization and estimation should be the primary tool for communicating the value of this work, and we have removed many of the hypothesis tests from the paper (see our point about removing the median splits analyses below). We now focus more on estimation and reporting the uncertainty in our estimates using intervals; an approach that we think fits well with the goals of Psychological Science's new reporting standards. We also removed the justification about robustness to outliers, and we instead try to motivate our Bayesian approach by emphasizing the value of including relevant prior information that constrain the parameter estimates and inform the Bayes Factors (BF). And while we agree that the sensitivity analyses show that the specification of the prior distribution is not critical for parameter estimation, this information is important for the BF values, which provide helpful information about the strength of the evidence. Finally, we think that the choice of a Bayesian approach is also justified by work showing that traditional frequentist statistical concepts (e.g., p-values and confidence intervals) are often interpreted as Bayes Factors and 95% Highest Density Intervals (e.g., Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J., 2014; Wasserstein, R. L., & Lazar, N. A., 2016).

> *It draws attention away from what is important. Graphics tell the story here (e.g. page 19, 20). On top of what we can see in these graphs, the Bayesian linear models only add complexity. Why commit to a linear model when a data description approach (with perhaps a nonparametric regression like LOWESS and nonparametric correlation like Kendall's tau) does not require such assumptions?*

We appreciate this point about reducing the complexity of our analysis approach, and in our revision we have tried to clarify just how straightforward it is. We chose a linear model because it is simple in terms of the number of parameters to estimate and provides an estimate of developmental change. And while we agree that there is value in using a nonparameteric approach (i.e., not assuming linear development), we were concerned that a LOWESS model would be difficult to interpret since this approach often requires lots of data to fit effectively and the interpretation can change depending on the complexity of the model (i.e., the value of the smoothing parameter). Thus, we felt that the linear models provided the best combination of simplicity and interpretability and chose to keep them in the revision.

> *In several places (p 9 - note that including two signers instead of 1 doesn't really help generalisation much, because you're stuck with N=2 instead of N=1 now ; and p 23) the authors mention the problem of generalisation. They not the problem is almost inherent*

*to what they study, which I find a compelling argument. But the statistics - particularly the hypothesis tests - uses *assume* that very generalisability. What population is being generalised to? It is not clear, and the authors know this. I think it is laudable to state possible problems with generalisability. I would also like to see statistic being treated in a bit more sophisticated manner: if the topic is interesting, and the sample useful sometimes inferential statistics are not necessary. Doing tests for the sake of generalising to some unknown hypothetical population, as opposed to simply presenting descriptive results for what they are, is not a good way to proceed, in my opinion.*
I think with an overhaul to remove the Bayesian linear modelling and inferential statistics, to focus on the descriptive statistics and linking it more strongly with what, graphically, we would expect from data from the LWL task, this would make for a much better paper.

We agree that we included several unnecessary hypothesis tests and have removed the majority of these tests from the revision. We now focus on estimation and reporting uncertainty around our estimates using the 95% HDIs. We also appreciate the point about motivating the use of inferential statistics. And while we agree that these results might not generalize to children learning sign language from inconsistent sign input, we do think that these results should generalize to our target population -- children who are learning a sign language under typical learning conditions. Moreover, studying this population (native ASL learners) allows researchers to understand the features of acquisition and processing that are language-general vs. modality-specific. Thus, we do think that there is a target population to generalize to, and as a result we chose to keep the planned hypothesis tests of developmental change and links to vocabulary in the revision.

Figure 4 needs work. On the x axis, a continuous variable is treated categorically, and bar plots with standard errors hide important aspects of the data. Each data point should be plotted. If the x variable can be treated continuously you can add in the group membership as a point shape. If one must treat the x variable as continuous, then try a violin plot with overlaid points.

Thank you for pointing out these issues. We changed the figure in several ways. First, we removed the median split by age in the children since this creates an arbitrary grouping variable based on the current sample. In the new analysis, children are compared to adults in order to provide information about the developmental endpoint relative to children's processing skills. Second, we split the children into hearing and deaf groups (a meaningful natural experiment) to show that these groups of ASL-learners (matched in age) perform similarly when allocating gaze during real-time ASL processing, despite having very different sensory experiences of the world. And third, we replaced the bar plots with violin plots (thank you for the suggestion) to show the full posterior distribution of, providing the reader with more information about the uncertainty around our point estimates of Accuracy and RT for both children and adults.