Supplementary materials for the paper

"Real-time lexical comprehension in young children learning American Sign Language"

In this document, we present four pieces of supplemental information. First, we provide details about the Bayesian models used to analyze the data. Second, we present a sensitivity analysis that provides evidence that the estimates of the associations between age/vocabulary and accuracy/reaction time (RT) are robust to different parameterizations of the prior distribution and different cutoffs for the analysis window. Third, we present the results of a parallel set of analyses using a non-Bayesian approach to show that these results are consistent regardless of choice of analytic framework. And fourth, we present two exploratory analyses measuring the effects of phonological overlap and iconicity on RT and accuracy. In both analyses, we did not see evidence that these factors changed the dynamics of eye movements during ASL processing.

Model Specifications

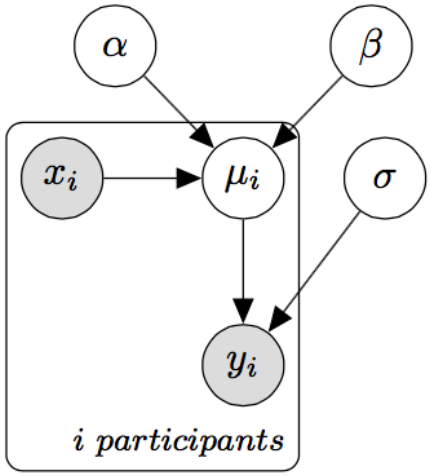
Our key analyses use Bayesian linear models to test our hypotheses of interest and to estimate the associations between age/vocabulary and RT/accuracy. Figure S1 (Accuracy) and S2 (RT) present graphical models that represent all of the data, parameters, and other variables of interest, and their dependencies. Latent parameters are shown as unshaded nodes while observed parameters and data are shown as shaded nodes. All models were fit using JAGS software (Plummer, 2003) and adapted from models in Kruschke (2014) and Lee and Wagenmakers (2014).

Accuracy

To test the association between age/vocabulary and accuracy we assume each participant's mean accuracy is drawn from a Gaussian distribution with a mean, μ, and a standard deviation, σ. The mean is a linear function of the intercept, α, which encodes the expected value of the outcome variable when the predictor is zero, and the slope, β, which encodes the expected change in the outcome with each unit change in the predictor (i.e., the strength of association).

For α and σ, we use vague priors on a standardized scale, allowing the model to consider a wide range of plausible values. Since the slope parameter β is critical to our hypothesis of a linear association, we chose to use an informed prior: that is, a truncated Gaussian distribution with a mean of zero and a standard deviation of one on a standardized scale. Centering the distribution at zero is conservative and places the highest prior probability on a null association, to reduce the chance that our model overfits the data. Truncating the prior encodes our directional hypothesis that accuracy should increase with age and larger vocabulary size. And using a standard deviation of one constrains the plausible slope values, thus making our alternative hypothesis more precise. We constrained the slope values based on previous research with children learning spoken language showing that the average gain in accuracy for one month of development between 18-24 months to be ~1.5% (Fernald, Zangl, Portillo, & Marchman,

2008).



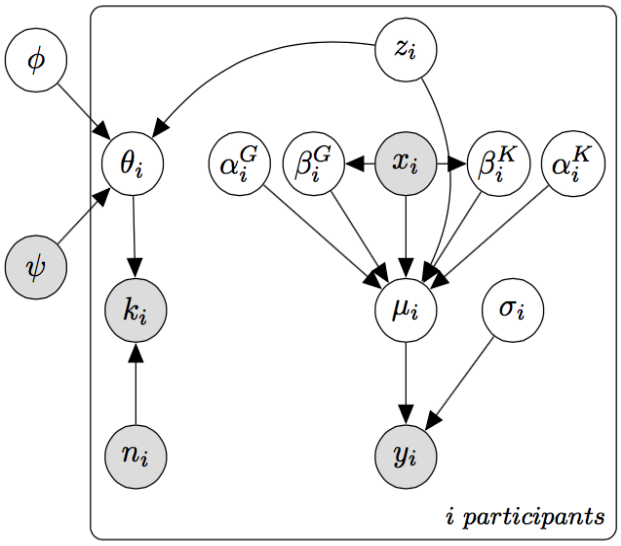
$$\alpha \sim \text{Gaussian}(0,\ 10)$$
$$\beta \sim \text{TruncatedGaussian}_{+}(0,\ 1)$$
$$\sigma \sim \text{Uniform}(1/1000,\ 1000)$$
$$\mu_i \leftarrow\ \alpha + \beta * x_i$$
$$y_i \sim\ \text{Gaussian}(\mu_i,\ \sigma)$$

Figure S1. Graphical model representation of the linear regression used to predict accuracy. The shaded nodes represent observed data (i.e., the i_{th} participant's age, vocabulary, and mean accuracy). Unshaded nodes represent latent parameters (i.e., the intercept and slope of the linear model).

Reaction Time

The use of RT as a processing measure is based on the assumption that the timing of a child's first shift reflects the speed of their incremental language comprehension. Yet, some children have a first shift that seems to be unassociated with this construct: their first shift behavior appears random. We quantify this possibility for each participant explicitly (i.e., the probability that the participant is a "guesser") and we create an analysis model where participants who were more likely to be guessers have less of an influence on the estimated relations between RT and age/vocabulary.

To quantify each participant's probability of guessing, we computed the proportion of signer-to-target (correct) and signer-to-distracter (incorrect) shifts for each child. We then used a



$$
\begin{aligned}
\alpha &\sim \text{Gaussian}(0,\ 10) \\
\beta &\sim \text{TruncatedGaussian\_}(0,\ 1) \\
\sigma_i &\sim \text{Uniform}(1/1000,\ 1000) \\
z_i &\sim \text{Bernoulli}(0.5) \\
\phi &\sim \text{Uniform}(0.5,\ 1) \\
\psi &\leftarrow 0.5 \\
\theta_i &\leftarrow \begin{cases} \phi & \text{if } z_i = 1 \\ \psi & \text{if } z_i = 0 \end{cases} \\
k_i &\sim \text{Binomial}(\theta_i,\ n_i) \\
\mu_i &\leftarrow \begin{cases} \alpha_i^K + \beta_i^K * x_i & \text{if } z_i = 1 \\ \alpha_i^G + \beta_i^G * x_i & \text{if } z_i = 0 \end{cases} \\
y_i &\sim \text{Gaussian}(\mu_i,\ \sigma_i)
\end{aligned}
$$

Figure S2. *Graphical model representation of the linear regression plus latent mixture model (i.e., guessing model). The model assumes that each individual participant's first shift is either the result of guessing or knowledge. And the latent indicator z_i determines whether the i_{th} participant is included in the linear regression estimating the association between age/vocabulary and RT.*

latent mixture model in which we assumed that the observed data, k_i, were generated by two processes (guessing and knowledge) that had different overall probabilities of success, with the "guessing group" having a probability of 50%, ψ, and the "knowledge" group having a probability greater than 50%, ϕ. The group membership of each participant is a latent indicator variable, z_i, inferred based on that participant's proportion of correct signer-to-target shifts relative to the overall proportion of correct shifts across all participants (see Lee & Wagenmakers (2014) for a detailed discussion of this modeling approach). We then used each participant's inferred group membership to determine whether they were included in the linear regression. In sum, the model allows participants to contribute to the estimated associations between age/vocabulary and RT *proportional* to our belief that they were guessing.

As in the Accuracy model, we use vague priors for α and σ on a standardized scale. We again use an informed prior for β, making our alternative hypothesis more precise. That is, we constrained the plausible slope values based on previous research with children learning spoken

language showing that the average gain in RT for one month of development between 18-24 months to be ~30ms (Fernald, Zangl, Portillo, & Marchman, 2008).

Sensitivity Analysis: Prior Distribution and Window Selection

We conducted a sensitivity analysis to show that our parameter estimates for the associations between accuracy/RT and age/vocabulary are robust to decisions about (a) the analysis window and (b) the specification of the prior distribution on the slope parameter. Specifically, we varied the parameterization of the standard deviation on the slope, allowing the model to consider a wider or narrower range of values to be plausible a priori. We also fit these different models to two additional analysis windows +/- 300 ms from the final analysis window: 600-2500 ms (the middle 90% of the RT distribution in our experiment).

Figure S3 shows the results of the sensitivity analysis, plotting the coefficient for the β parameter in each model for the three different analysis windows for each specification of the prior. All models show similar coefficient values, suggesting that inferences about the parameters are not sensitive to the exact form of the priors. Table S1 shows the Bayes Factors for all models across three analysis windows and fit using four different vales for the slope prior. The Bayes Factor only drops below 3 when the prior distribution is quite broad (standard deviation of 3.2) and only for the longest analysis window (600-2800 ms). In sum, the strength of evidence for a linear association is robust to the choice of analysis window and prior specification.

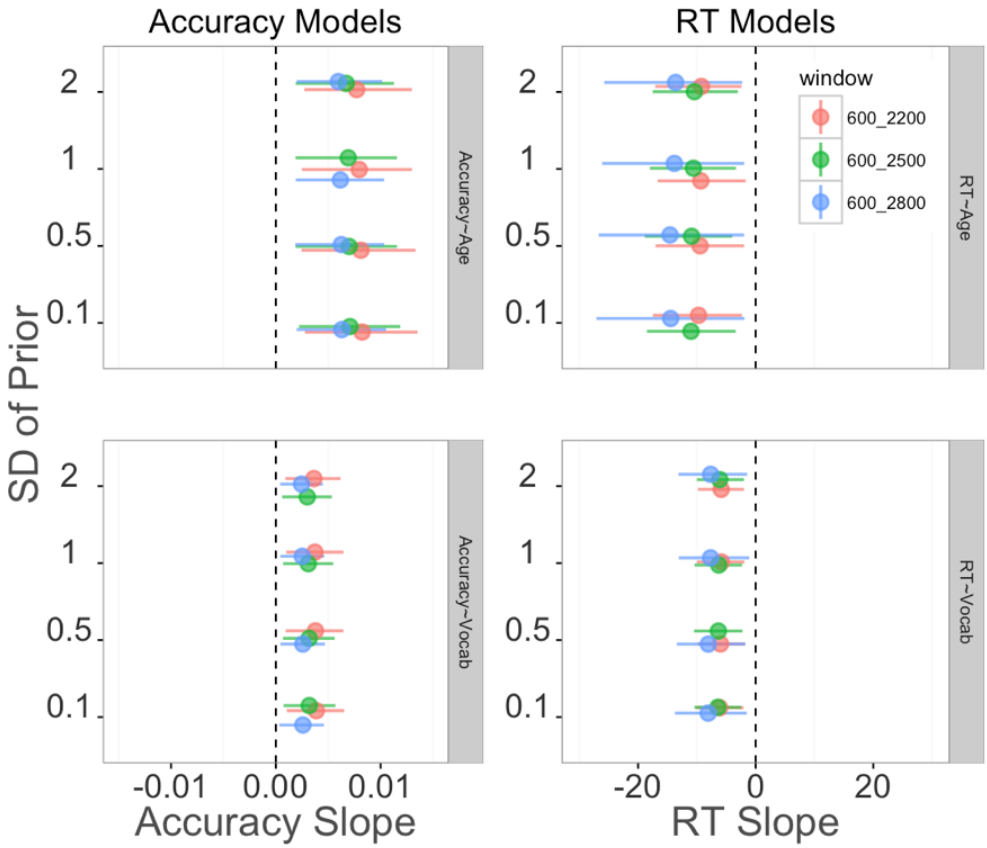


Figure S3. *Coefficient plot for the slope parameter, β, for four different parameterizations of the prior and for three different analysis windows. Each panel shows a different model. Each point represents a β coefficient measuring the strength of association between the two variables. Error bars are 95% HDIs around the coefficient. Color represents the three different analysis windows.*

Analysis window	SD Slope	Acc~Age	Acc~Vocab	RT~Age	RT~Vocab
600 – 2200 ms	3.2	6.2	3.7	2.4	4.1
	1.4	14.1	5.5	3.5	8.6
	1.0	19.4	8.9	5.0	9.2
	0.7	22.7	11.6	7.8	17.0
600 – 2500 ms	3.2	11.0	2.3	5.6	6.1
	1.4	9.7	4.0	13.8	10.5
	1.0	12.8	6.8	12.5	18.2
	0.7	15.6	6.8	17.9	20.7
600 – 2800 ms	3.2	6.0	1.1	1.2	1.4
	1.4	10.7	2.6	3.5	4.7
	1.0	13.5	4.0	3.7	4.0
	0.7	15.2	4.6	5.5	5.6

Table S1. *Bayes Factors for all four linear models fit to three different analysis windows using four different parameterizations of the prior distribution for the slope parameter β.*

Parallel set of non-Bayesian analyses

First, we compare Accuracy and RT of native hearing and deaf signers using a Welch Two Sample t-test and do not find evidence that these groups are different (Accuracy: $t(28) = 0.75$, $p = 0.45$, 95% CI on the difference in means [-0.07, 0.14]; RT: $t(28) = 0.75$, $p = 0.46$, 95% CI on the difference in means [-125.47 ms, 264.99 ms].

Second, we test whether children and adults tend to generate saccades away from the central signer prior to the offset of the target sign. To do this, we use a One Sample t-test with a null hypothesis that the true mean is not equal to 1, and we find evidence against this null (Children: $M = 0.88$, $t(28) = -2.92$, $p = 0.007$, 95% CI [0.79, 0.96]; Adults: $M = 0.51$, $t(15) = -6.87$, $p < 0.001$, 95% CI [0.35, 0.65])

Third, we fit the four linear models using MLE to estimate the relations between the processing measures on the VLP task (Accuracy/RT) and age/vocabulary. We follow recommendations from Barr (2008) and use a logistic transform to convert the proportion accuracy scores to a scale more suitable for the linear model.

Model specification	β value	std. error	t-statistic	p-value
logit(accuracy) ~ age + hearing status	0.003	0.012	2.59	0.008
RT ~ age + hearing status	-10.05	4.62	-2.17	0.019
logit(accuracy) ~ vocabulary + hearing status	0.002	0.006	2.27	0.015
RT ~ vocabulary + hearing status	-6.34	2.18	-2.91	0.003

Table S2. *Results for the four linear models fit using MLE. All p-values are one-sided to reflect our directional hypotheses about the VLP measures improving over development.*

Analyses of phonological overlap and iconicity

First, we analyzed whether phonological overlap of our item-pairs might have influenced adults and children's RTs and accuracy. Signs that are higher in phonological overlap might have been more difficult to process because they are more confusable. Here, phonological overlap is

quantified as the number of features (e.g., Selected Fingers, Major Location, Movement, Sign Type) that both signs shared. Values were taken from a recently created database (ASL-LEX) of lexical and phonological properties of nearly 1,000 signs of American Sign Language (Caselli et al., 2017). Our item-pairs varied in degree of overlap from 1-4 features. We did not see evidence that degree of phonological overlap influenced either processing measure in the VLP task.

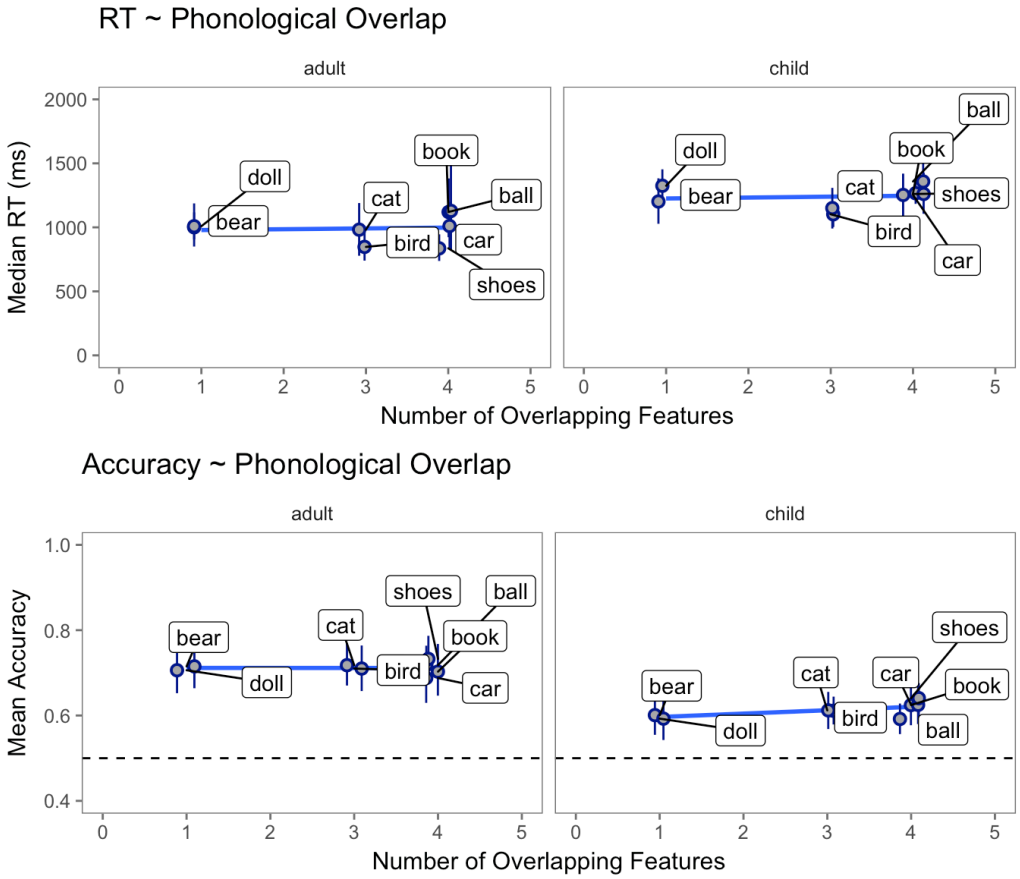


Figure S4. *Scatterplot of the association between degree of phonological overlap and RT (top row) and accuracy (bottom row) for both adults (left column) and children (right column). The blue line represents a linear model fit.*

Next, we performed a parallel analysis, exploring whether the iconicity of our signs might have influenced adults and children's RT and accuracy. It is possible that highly iconic signs might be easier to process because of the visual similarity to the target object. Again, we

used ASL-LEX to quantify the iconicity of our signs. To generate these values, native signers were asked to explicitly rate the iconicity of each sign on a scale of 1-7, with 1 being not iconic at all and 7 being very iconic. Similar to the phonological overlap analysis, we did see evidence that degree of iconicity influenced either processing measure for either age group in the VLP task.

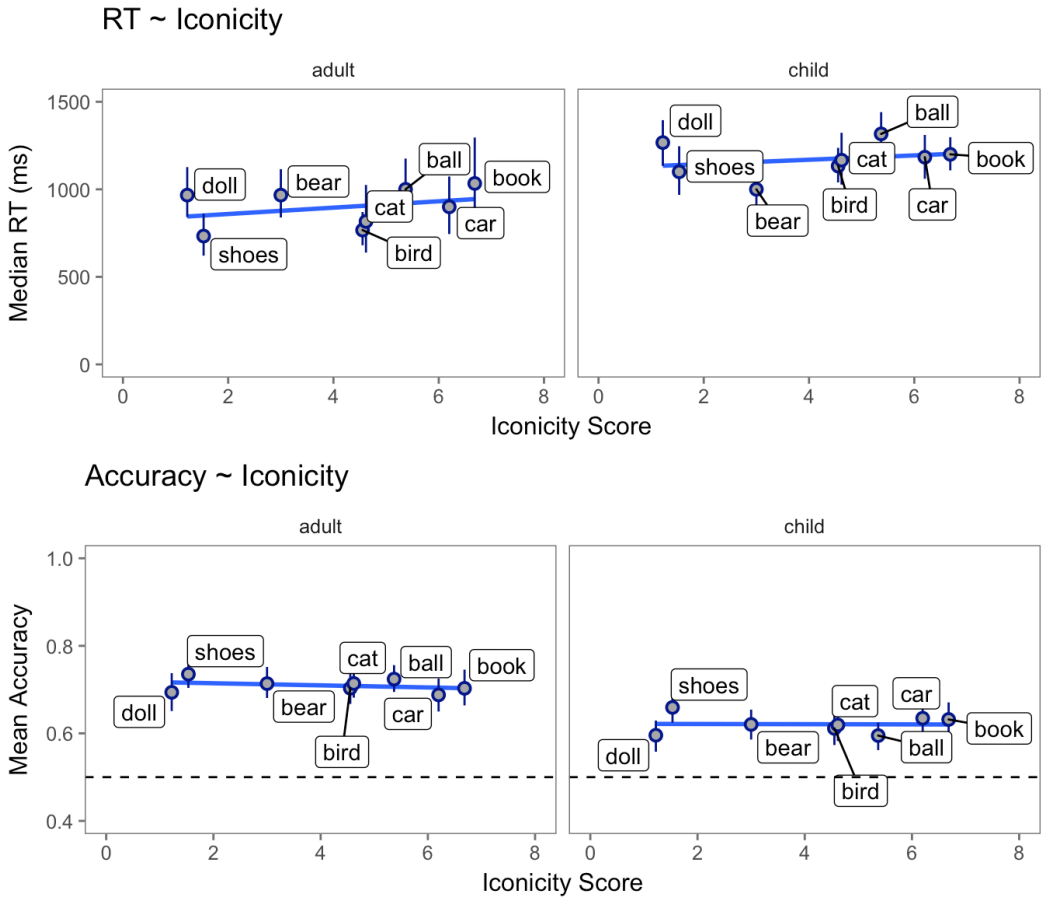


Figure S5. *Scatterplot of the association between degree of iconicity and RT (top row) and accuracy (bottom row) for both adults (left column) and children (right column). The blue line represents a linear model fit.*

References

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457-474.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. *Behavior research methods*, *49*(2), 784-801.

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. In Sekerina, I.A., Fernandez, E.M., & Clahsen, H. (Eds.). *Developmental psycholinguistics: On-line methods in children's language processing*, 113-132.

Kruschke, J. (2014). Doing Bayesian data analysis: A tutorial with r, jags, and stan. Academic Press.

Lee, M. D., & Wagenmakers, E.J. (2014). Bayesian cognitive modeling: A practical course. Cambridge University Press.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing* (Vol. 124, p. 125). Wien, Austria: Technische Universit at Wien.