**Supporting Online Material**

**Method**

Participants

Participants were 16 deaf and 13 hearing children with native exposure to ASL (17 females, 12 males, = 28.5 months, range = 16-53 months) and 19 fluent adult signers. Our sample size was determined by our success over the 2-year period of an NIDCD R21 grant in recruiting and testing children and adults who were native ASL users. . An additional 20 child participants were tested but not included in the analyses because they were not exposed to ASL from birth (*n* = 14), or they did not complete the real-time language assessment (*n* = 6).

Children learning ASL from birth from a native signer are a difficult population to recruit, given that approximately 95% of deaf children are born to hearing parents with little or no prior exposure to a signed language (Mitchell & Karchmer, 2004). The majority of child participants were recruited through a center-based child education program in which ASL was the mode of instruction. All children were exposed to ASL from birth through extensive interaction with at least one fluent ASL caregiver, and they currently used ASL as their primary mode of communication at home. Adult participants were all fluent signers who reported using ASL as their primary method of communication. All participants who met the inclusionary criteria and who had complete data were analyzed

Data processing

Prior to coding, all trials for child participants were pre-screened to exclude those few with parental interference on a trial-by-trial basis.

Computing target sign onset

In studies of spoken language processing, target word onset is typically identified as the first moment in the auditory signal when there acoustic evidence of the target word. However, in signed languages like ASL, phonological information is presented in several parts of the visual signal simultaneously – for example, in both the hands and face of the signer - making it difficult to precisely determine the beginning of the target sign. In the VLP task, this problem is simplified because pictures are presented in yoked pairs; thus on each trial, target sign onset is always determined in the context of a particular distracter object, the name of which does not overlap phonologically with the target object. Thus, target sign onset can be defined as the earliest point in the signed sentence when the two pictures can be reliably discriminated.

Here, we took an empirical approach to defining target sign onset. For each of 28 sign tokens, we created six videos, each showing a different amount of the target sign, ranging from three frames before to three frames after the noun onset determined based on our initial consensus judgments. We then asked 10 fluent adult signers unfamiliar with the stimuli to watch these videos while viewing the same picture pairs as in the VLP task. On each trial, they made forced-choice decisions indicating which image was signed in the video, yielding a proportion correct target identification for each video. For each sign token, final target noun onsets were identified as the earliest point in the signed sentence at which adults discriminated the pictures with 100% agreement.

Measures of processing efficiency

*Reaction Time.* In the VLP task, four different types of responses are possible on a given trial: (1) signer-to-target object shift, (2) signer-to-distracter object shift, (3) signer-to-away shift, and (4) no shift. Since children varied in signer-to-target shifts, mean RTs were based on different numbers of trials across participants (*M* = 12.7 trials, range = 3-25).

*Accuracy.* Although all children and response types were included in the computation of accuracy, the number of trials contributing to the analysis varied across participants (*M* = 19.1).

**Results**

Model specifications

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. Previous work using the Looking-while-Listening paradigm could not easily compute these values, since the task did not include a center fixation point. We then used a latent mixture model in which we assumed that the observed data (children’s initial shifts away from the signer) were generated by two processes (guessing and knowledge) that had different overall probabilities of success, with the “guessing group” having a probability of 50% and the “knowledge” group having a probability > 50%. The group membership of each participant was a latent variable 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 [2013] for a detailed discussion of this modeling approach). We then used each participant’s inferred group membership to weight participants *proportional* to our belief that they were guessing[[1]](#footnote-1). It is important to point out that we use this approach only in the analysis of RT, given our assumption that “guessing behavior” is integral to our measure of children’s mean accuracy in the VLP task, but not to our measure of mean RT which is based on correct trials.

In all of the Bayesian linear models[[2]](#footnote-2), we assume that each outcome variable (mean accuracies and RTs for each participant) is drawn from a Gaussian distribution with a mean, μ, and a standard deviation, σ. The mean is generated by a linear function consisting of an intercept term, α, which encodes the expected value of the outcome variable when the predictor is zero, and a slope term, β, which encodes the expected change in the outcome with each unit change in the predictor (i.e., the strength of association). We use vague priors for the intercept and the standard deviation, allowing the model to consider a wide range of plausible values.

We chose to use informative priors for the slope parameters in each model. Specifically, we used a truncated Gaussian distribution with a mean of zero and a standard deviation of one. Centering the distribution at zero is conservative and places the highest prior probability on a null association. By truncating the prior, we encoded our directional hypotheses for the associations between processing skills and age/vocabulary (i.e., that we predict that these relations should be null or improve with increasing age and larger vocabulary size). Finally, to constrain the range of plausible slope values in our model, we used previous research on the development of real-time language comprehension in children learning spoken language showing that the average gain for one month of development between 18-24 months in accuracy is ~0.016 and for RT is ~50 milliseconds (Fernald et al., 2008).

It is important to point out that our use of an informative prior does not affect parameter estimation. That is, if we substitute vague prior distributions, estimates of the strength of the associations between age/vocab and accuracy/RT are unchanged since there are enough data to overwhelm the uninformative priors. However, the use of an uninformative prior does directly influence Bayesian model comparison, which we use to quantify the strength of evidence for our linear models (i.e., the Bayes Factor[[3]](#footnote-3)). Intuitively, the use of an uninformative prior allows our models to predict any slope parameter, creating a situation where only a small amount of the prior probability is placed on a null relationship (i.e., where the slope is zero). Thus, the Bayes Factor, which is computed by taking the ratio of the prior and the posterior density when the slope is zero, is likely to show a preference for the null model, even when the data appear inconsistent with it (see Lee & Wagenmakers [2013]). The use of informative priors is an area of active debate in Bayesian statistical methods, but it has become a focus of recent work in Bayesian cognitive modeling (Lee & Vanpaemel, submitted).

Categorical models of target looking

Moreover, when we modeled mean target looking as a function of age group, the 95% HDI for each group did not include the value of 0.5 (younger:β= 0.61, 95% HDI [0.56, 0.67]; older:β= 0.70, 95% HDI [0.65, 0.76]; adults: β=0.83, 95% HDI [0.78, 0.88]), providing evidence that, as a group, the lower bound of the mean estimates for accuracy for

Sensitivity analysis

Window selection analysis

1. Four children (ages: 18, 20, 22, and 25 months) had high posterior probability mass on guessing: posterior probabilities of 0.89, 0.86, 0.82, and 0.77, respectively (mean proportion signer-to-target scores for these participants were: 0.55, 0.46, 0.38, 0.33). [↑](#footnote-ref-1)
2. Models with categorical predictors were implemented in STAN (Stan Development Team, 2016). Models with continuous predictors were implemented in JAGS (Plummer, 2003). [↑](#footnote-ref-2)
3. The Bayes Factor can be interpreted as a measure of the relative strength of evidence one model (M1) over another model (M2): e.g., a of 5 means that the data are 5 times more likely given M1. [↑](#footnote-ref-3)