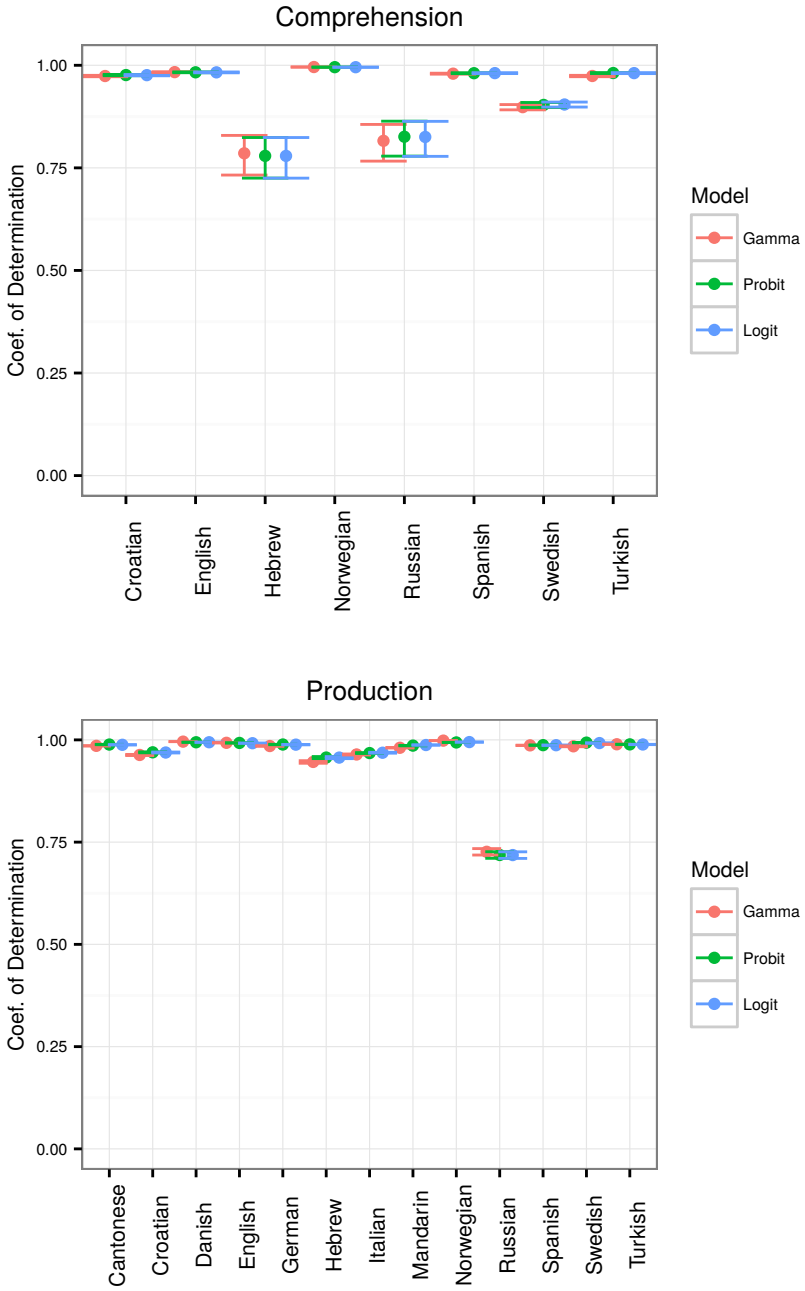

Figure S1: Model comparison of the Logit, Probit and Gamma models when trained on the full learning curve. Across words and languages, the correlations between observed data and model predictions are close to 1.

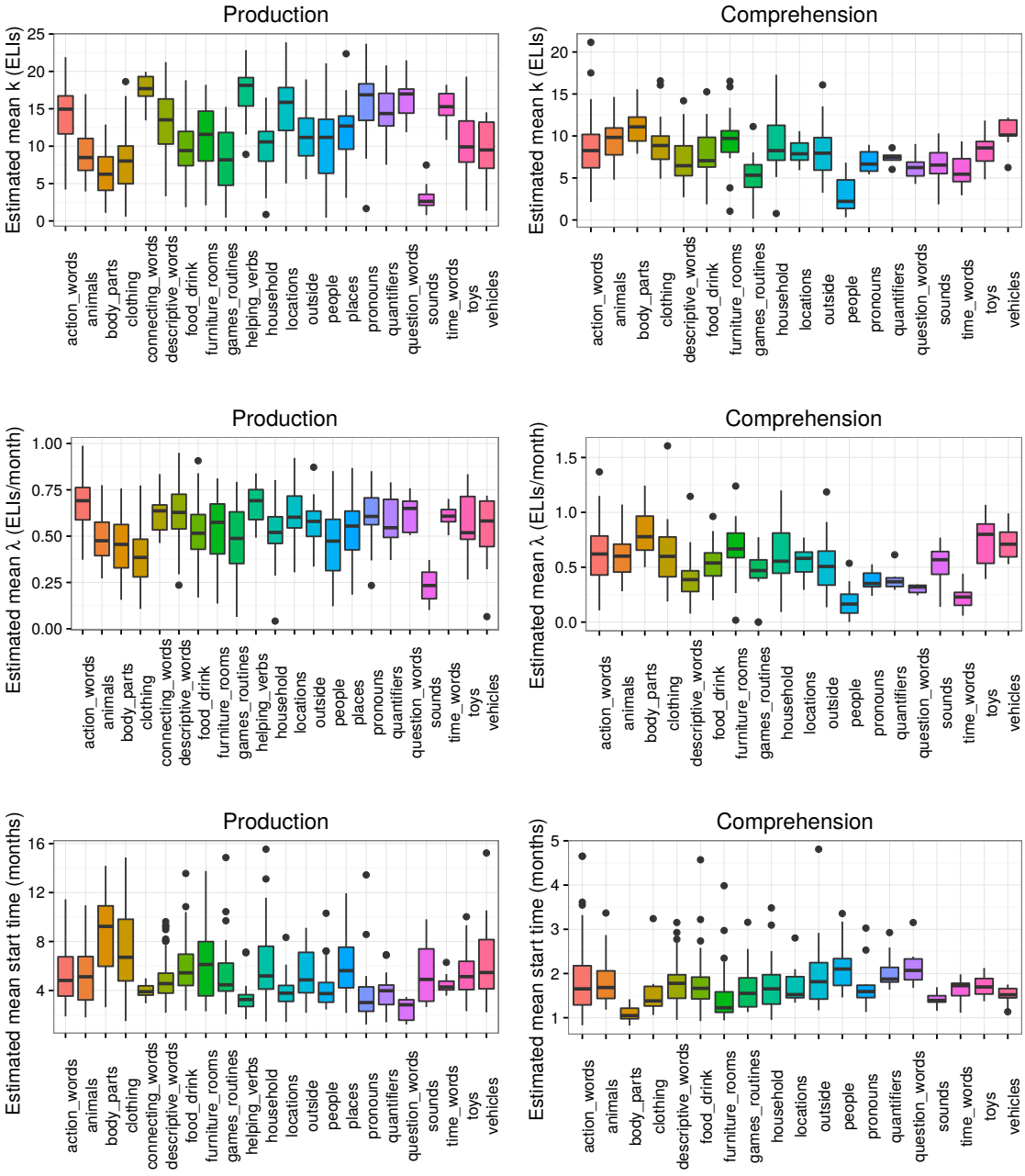

Figure S2: Box plots of the mean k, λ and s values measured for English words split by MCDI semantic category.

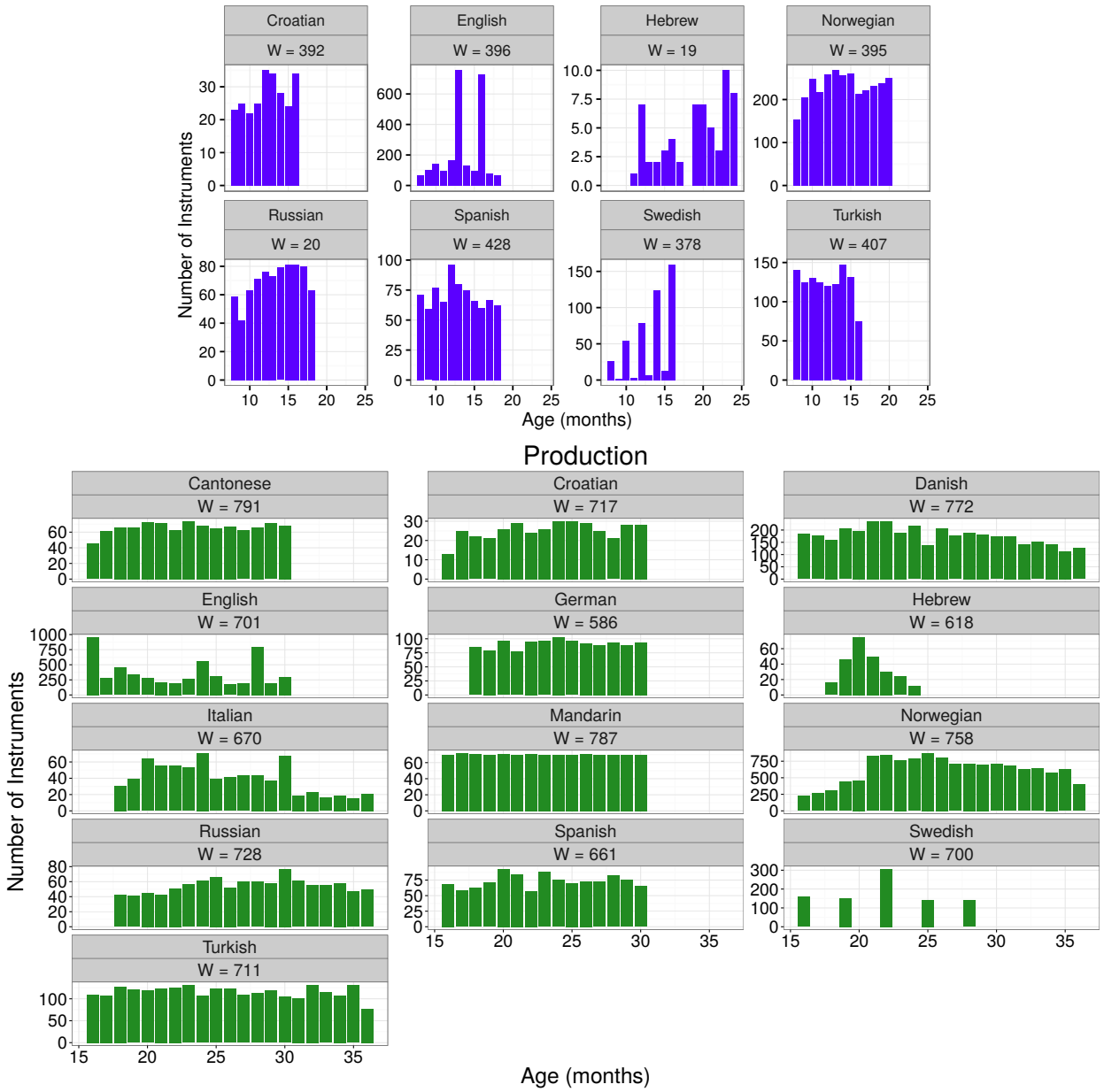

Figure S3: Number of completed MCDIs at each age for each language and words for each instrument. Note the y-axes differ in each panel.

A Model Justification

Hidaka (2013) conducted a model comparison of three different generative models for the AoA distributions: a rate-change learning model (i.e., a Weibull model), a cumulative learning model (i.e., a Gamma model), and a cumulative-and-rate-change learning model (i.e., a Weibull-Gamma model). In the rate-change model, a learner only requires a single ELI and each month the initial probability of observing an ELI, λ, changes (presumably increases) polynomially, with an exponent of δ. The cumulative learning model is the gamma model we chose to implement (without a start time parameter), i.e., a learner requires k ELIs to learn a word and ELIs come stochastically but at a fixed rate, λ ELIs/month. In the cumulative-and-rate-change model, a learner requires k effective learning instances to learn a word and these instances have a base rate λ which changes by a power of δ each month.

The cumulative distribution function of these three models describes the probability of a child having learned a word as a function of age a. The equations for these three models follow:

Weibull Model

$$F(a;\lambda,\delta) = \gamma(1,(\lambda \cdot a)^{\delta}) \tag{4}$$

Gamma Model

$$F(a;k,\lambda) = \frac{\gamma(k,\lambda \cdot a)}{\Gamma(k)} \tag{5}$$

Weibull-Gamma Model

$$F(a;k,\lambda,\delta) = \frac{\gamma(k,(\lambda \cdot a)^{\delta})}{\Gamma(k)} \tag{6}$$

where $\gamma(k,b)$ is the lower incomplete gamma function,

$$\gamma(k,b) = \int_0^b t^{k-1}e^{-t}dt. \tag{7}$$

and $\Gamma(k)$ is the gamma function:

$$\gamma(k) = \int_0^\infty t^{k-1}e^{-t}dt. \tag{8}$$

The parameter k is interpreted the same as above—i.e., the number of ELIs required for learning. The parameter b corresponds to the expected number of instances observed at that time.

Hidaka (2013) fit these models for 652 productive vocabulary words in the MCDI. He found that when aggregating over words, the cumulative-and-rate-change model has the best fit as measured by Bayesian Inference Criterion (BIC). However, when he looked at each word individually and compared the BICs for the different models, he found that the cumulative model fits best for 50% of the words. This means that the cumulative-and-rate-change model is a good overall model of early word learning[5], but the generative process that best captures how individual words are learned is the cumulative (Gamma) model.

We chose to use and extend the cumulative (Gamma) model for our purposes for two reasons. First, the majority of the individual words fit by Hidaka (2013) were best fit by the Gamma model, making the option of choice for capturing individual word curves. Second, the Gamma model has a more straightforward interpretation than the rate change models. While the k parameter retains the same units (number of ELIs) across all models[6], the interpretation of the parameter b differs across models. For the gamma model, the unit for b—i.e., expected ELIs, does not change across time. However, in the rate change model, b is raised to an exponent, giving it a much less clear interpretation. It effectively gives rise to some polynomial of time, but not one which is to our knowledge motivated by independent considerations. These two factors lead us to build off the gamma model rather than the rate change model, even though the latter fits better in one analysis.

Lastly, we can directly compare the performance of cumulative-and-rate-change models and our model in predicting our data. These results show that the model we focus on, a Gamma model with start time, does a much better job than others in predicting learning curves. Supplementary Figure S4 displays the

[5]This is not too surprising considering both the cumulative (Gamma) model and the rate-change (Weibul) model are special cases of the cumulative-and-rate-change model.

[6]In a Weibull distribution $k = 1$ ELI

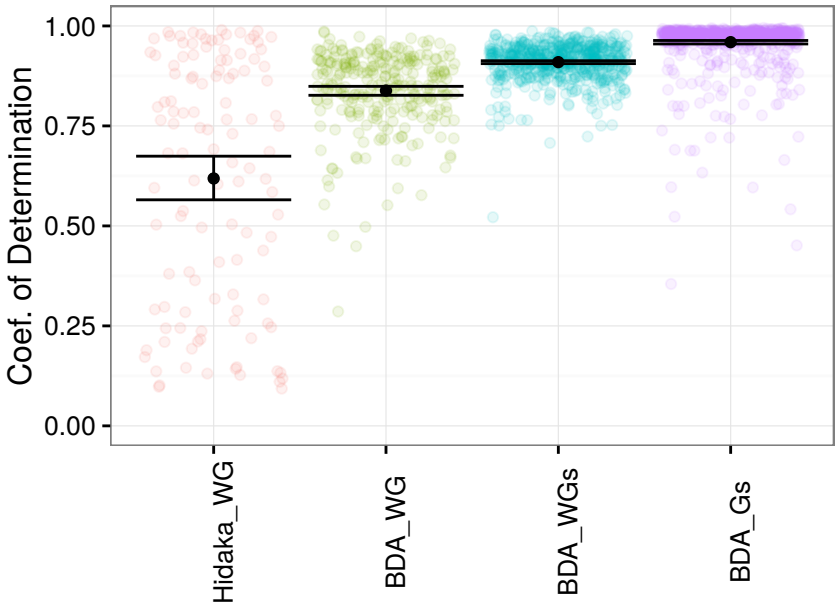

Figure S4: Model Comparison of Hidaka's Weibull-Gamma (Hidaka_WG), a BDA Weibull-Gamma (BDA_WG), a BDA Weibull-Gamma with a start time (BDA_WGs), and our Gamma model, which includes a start time (BDA_Gs). The low contrast points represent the coefficient of determination for individual words. The points represent the mean and their error bar's represent bootstrapped 95% confidence intervals.

coefficients of determination (across learning curves) of individual English words[7]. This figure contains four models: Hidaka (2013)'s estimated parameters for the Weibull-Gamma (Hidaka_WG), a Bayesian Data Analysis[8] (BDA) Weibull-Gamma model (BDA_WG), a BDA Weibull-Gamma with a start time parameter (BDA_WGs), and a Gamma model which includes a start time (BDA_Gs) (the primary one in our analysis).

Comparing Hidaka's Weibull-Gamma to the BDA_WG, we see that the BDA_WG provides a slightly better fit to the Wordbank MCDI data. Unfortunately we cannot distinguish if this might be due to the increased data amounts the parameters were estimated from or the robustness of Bayesian Data Analysis as a method. More interestingly, the inclusion of a start time parameter to the BDA_WG significantly increases the model fit to the data. Nonetheless, the Gamma model (BDA_Gs) still outperforms all of the cumulative-and-rate-change models. In other words, on these data sets, the model with a start time and no rate change provides the best fit.

However, given the nearness in fit between the BDA_WGs and our model, we compared the parameter values to see if the rate change parameter significantly differed from 1, which would suggest very little to no change in rate. We find that for 98% of the words, the rate change parameter is not significantly different than 1. This explains why BDA_Gs, which has this parameter set to 1, can perform so well.

To summarize, these results justify the use of a gamma model with start time in our primary analyses. However, it is important to remember that other age ranges or data sets may necessitate models with different probabilistic assumptions.

[7]Code and parameter estimates available on lab website.

[8]For the Bayesian Data Analysis models, the same inference procedure was used as in the main paper. The prior on the rate change parameter was $\delta \sim Gamma(2.25, 1.25)$. For the BDA_WG model, 319 of the 797 English MCDI words converged and were used in the analysis. Although the number of words successfully fit under this model seems low, Hidaka's WG parameter estimates only yield learning curves for 117 words. The WG parameters for the rest of the words Hidaka fit describe learning curves that are 1 or 0 over the window of data. For the BDA_WGs, 698 of the 797 words converged and were used in the analysis.

B Relaxing Our Data Analysis Assumption

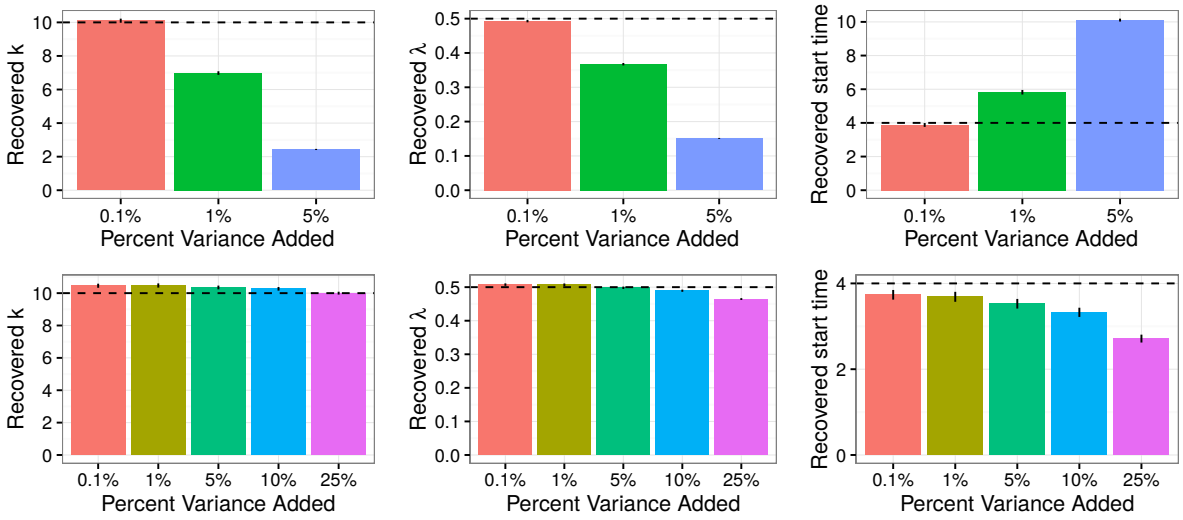

Figure S5: Recovered parameters estimates for simulations with varying percent model-internal noise (top row) and model-external noise (bottom row). The dashed line represents the generating parameter value. Point ranges reflects 95% bootstrapped confidence interval.

In our analysis, we make the data analysis assumption that the parameter values we infer will be the same across children; however, it is widely acknowledged that children implement different strategies for language learning (Brown, 1973). To examine how the model works when our assumption is violated, we simulated data with two different types of noise: model-internal noise—i.e, noise in the parameter values, and model-external noise. Model-internal noise might reflect individual differences in the learning process. Whereas, model-external noise might reflect things like measurement error.

We simulate data with model-internal noise by sampling parameters as follows:

$$\begin{aligned} k &\sim N(10, \sqrt{10v}) \\ \lambda &\sim N(0.5, \sqrt{0.5v}) \\ s &\sim N(4, \sqrt{4v}) \\ AoA &\sim \Gamma(k, \lambda) + s \end{aligned} \tag{9}$$

where v is the percent noise. To assess internal noise, we added either 0.1%, 1% and 5% noise. As the percent of internal noise increases, the shape of the AoA distribution to be fit changes significantly. For example, adding 1% internal noise increases the standard deviation of the AoA distribution by one month; whereas, adding 5% internal noise increases the standard deviation of the AoA distribution by at least 20 months. For each percent noise, we simulated age of acquisition data for 1000 children. We binned the simulated data across the age range of $15 - 36$ and ran the model on the binned data. We repeated this process 1000 times.

We expect that the recovered parameters from the model runs should be similar to the generating parameters. The results are shown in the top row of Supplementary Figure S5. First, note that with only 0.1% model-internal noise added, the recovered parameters are virtually the same as the generating parameters. Second, we find that under a reasonable percentage of added model-internal noise, the model recovers parameter values on the same order of magnitude as the generating parameters, suggesting that model-internal noise has a small effect on the order of magnitude of the parameter values. Lastly, we find that as the percent of model-internal noise increases, the recovered parameters for k and λ are under-estimated and the recovered parameter for s is over-estimated.

Given that the data in Wordbank (Frank et al., in press) , like all data, is inherently noisy, these simulations would suggest that our estimates for k and λ should be interpreted as lower bounds and our estimates for s should be interpreted as an upper bound. In effect, the presence of model-internal noise underestimates the contribution of data-driven processes to word learning and over-estimates the contribution of maturational processes. Despite this, we still find that the majority of the variance in early word learning can be explained by the simplest data-driven processes, i.e., waiting for data.

We simulate data with model-external noise by sampling ages of acquisition as follows:

$$\begin{aligned} a &\sim \Gamma(10, 0.5) + 4 \\ AoA &\sim N(a, \sqrt{av}) \end{aligned} \tag{10}$$

To assess external noise, we added either 0.1%, 1%, 5%, 10% or 25% noise. As can be seen in the bottom row of Supplementary Figure S5, We find that the model is remarkably robust at recovering the generating parameters when model-external noise is added.