Supplemental Material S1. Technical details on integrated weighted intelligibility response for multiword utterances.

Let L_i be the length of the longest utterance(s) achieved by child i, so that $2 \leq L_i \leq 7$ for each child. Let R_{ik} be the intelligibility score (i.e., the average across the two final listeners) for utterances of length k for child i, where R_{ik} is missing for $k > L_i$. Finally, let X_i be the child's age.

Problem. The challenge is that we do not want to analyze and report separately the intelligibility values R_{ik} for each utterance length k because (a) it would result in too many results, each with relatively weak information, getting weaker for larger k, and (b) the non-missing intelligibility values for larger k are only available for the children with higher levels of development.

Imputation and averaging. Here we generate a weighted average of utterance length-specific intelligibility values after imputing the missing values.

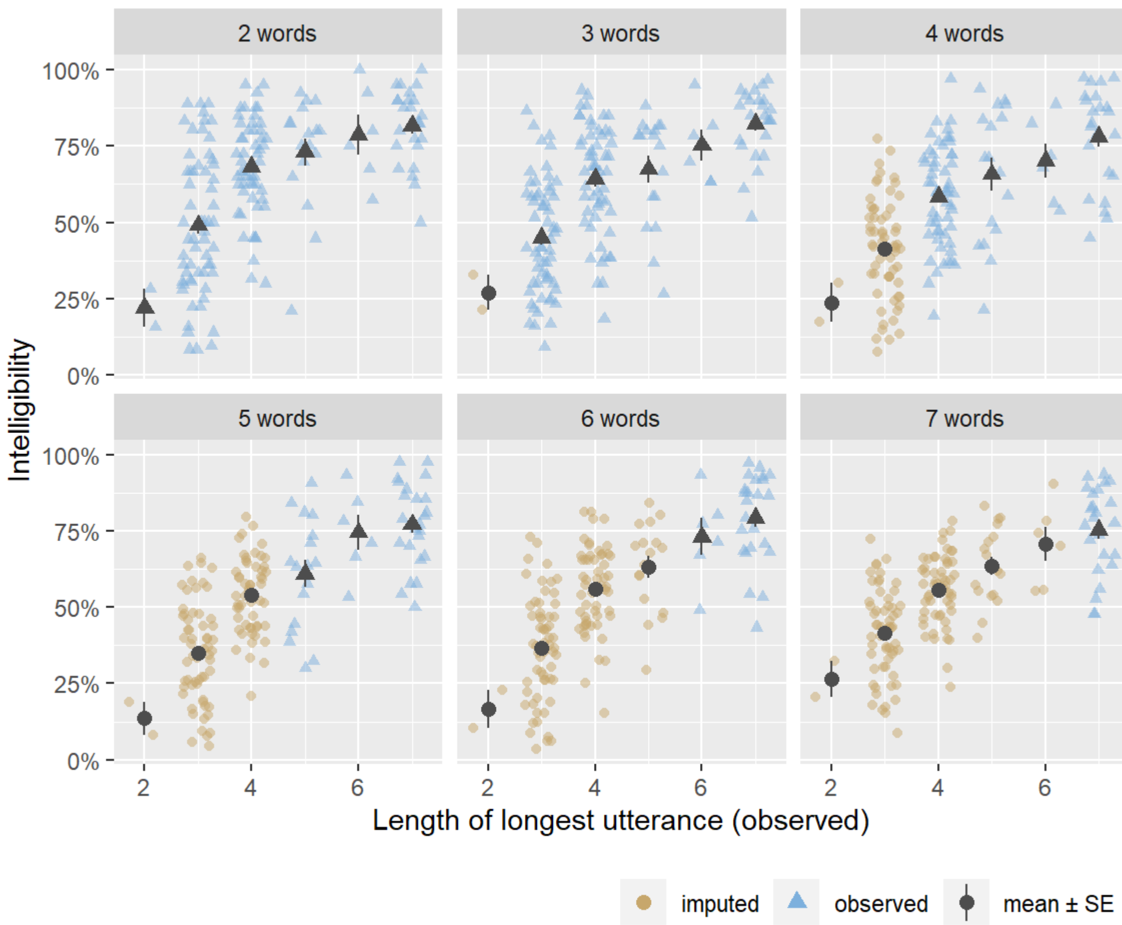
1. Using only the data for which $L_i = 7$, fit a regression model for score R_{i7} as a linear function of $R_{i2}, \ldots, R_{i6}$. Save the regression coefficients.
2. Sequentially fit 4 more regression models (for $k = 3, 4, 5, 6$) for R_{ik} as functions of $R_{i2}, \ldots, R_{i,k-1}, L_i$, noting the inclusion of L_i in this model). For each model fit, only use the data for which for $L_i \geq k$. Save the regression coefficients for each model.
3. Using the models, for all missing values of R_{ik} (i.e., when $L_i > k$), predict (impute) R_{ik} with the regression models. Call the new values $\tilde{R}_{ik}$. (These are the true R_{ik}'s when available or the imputed ones when not). Start from $k = 3$ and work the way up so in each case k, either true or imputed values $\tilde{R}_{ik}$ are used to predict the next level up.
4. Now, using a 2 df natural spline in age (X_i), fit an ordinal logistic regression model for L_i. From this model, obtain, for $k = 3, 4, 5, 6, 7$, probability $\tilde{\pi}_{ik}$ as a function of X_i that $L_i \geq k$. These probabilities will be decreasing in k. Note that $\tilde{\pi}_{i2} = 1$. Now normalize these values by computing $\pi_{ik} = \frac{\tilde{\pi}_{ik}}{\sum_{k'=2}^{7} \tilde{\pi}_{ik'}}$ so that they sum to 1.
5. Finally compute a weighted average Y_i of $R_{i2}, \ldots, R_{i7}$ using weights $\pi_{i2}, \ldots, \pi_{i7}$. Call this weighted average $\tilde{R}_i$. These values serve as our *integrated weighted intelligibility response* for multiword utterances.

Discussion. Our approach is very algorithmic; further methodological investigation would fully specify a statistical model for latent intelligibility which would (a) give rise to the observed (manifest) L_i values as well as to the values of the observed instances of R_{ik}.

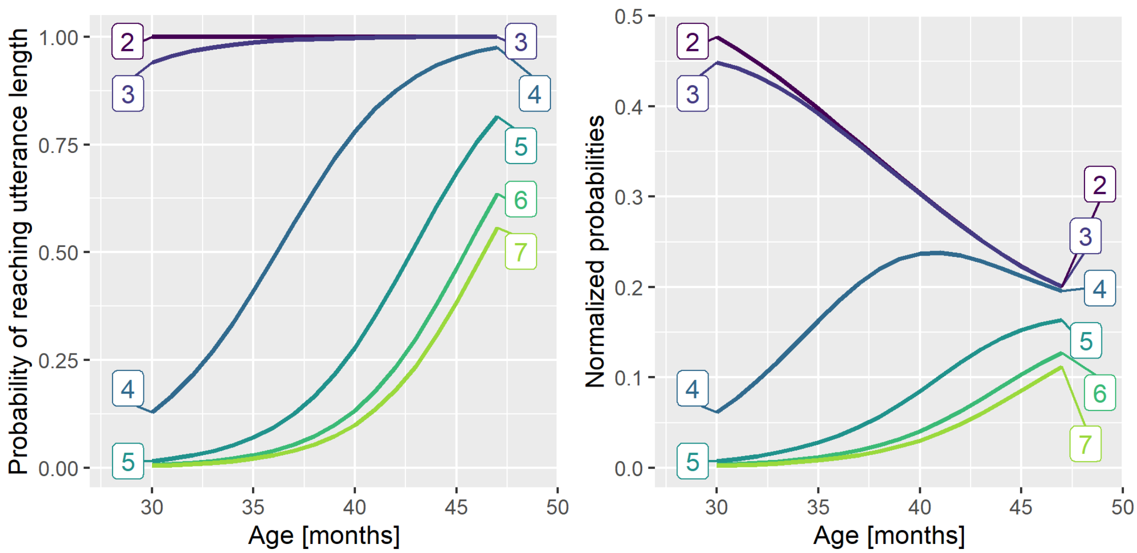
Coefficients used for imputation (1, 2).

Outcome	*Predictor*	*Estimate*	*SE*	*t*	*p*	*N*	*Adj. R^2*
3-wd	(Intercept)	−.07	.04	−1.70	.091	162	.70
	1-wd	.42	.10	4.27	< .001		
	2-wd	.45	.07	6.69	< .001		
	Length of longest utt.	.03	.01	4.04	< .001		
4-wd	(Intercept)	−.13	.07	−1.79	.077	102	.58
	1-wd	.45	.14	3.28	.001		
	2-wd	.21	.11	1.88	.063		
	3-wd	.36	.10	3.46	< .001		
	Length of longest utt.	.02	.01	1.59	.115		
5-wd	(Intercept)	−.15	.14	−1.04	.304	47	.61
	1-wd	.06	.21	0.27	.786		
	2-wd	.33	.15	2.25	.030		
	3-wd	.20	.16	1.25	.219		
	4-wd	.29	.13	2.26	.029		
	Length of longest utt.	.03	.02	1.64	.108		
6-wd	(Intercept)	−.13	.35	−0.37	.716	30	.49
	1-wd	.24	.24	1.01	.322		
	2-wd	.04	.23	0.17	.870		
	3-wd	.33	.27	1.23	.232		
	4-wd	.07	.18	0.41	.687		
	5-wd	.39	.22	1.79	.086		
	Length of longest utt.	.01	.05	0.24	.810		
7-wd	(Intercept)	.09	.14	0.60	.555	24	.70
	1-wd	.24	.20	1.18	.253		
	2-wd	−.06	.19	−0.34	.741		
	3-wd	−.35	.25	−1.40	.180		
	4-wd	.19	.14	1.30	.210		
	5-wd	.15	.19	0.78	.445		
	6-wd	.71	.17	4.10	< .001		

Imputation results (3). The following figure shows the intelligibility scores for each utterance length (panels) by length of lonest utterance (*x* axis). The blue triangles are observed and gold circles are imputed.



Weighting of utterance lengths by age (4). The following figure shows the probability of reaching each utterance length as a function of age (left). These probabilities were normalized and used as weights for computing the overall multiword intelligibility average (right). Thus, at 30 months, over 90% of the weighting comes from the 2- and 3-word utterances but by 47 months, 40% of the weighting comes from 2- and 3-word utterances.



Computing the final weighted average (5). The following figure compares multiword intelligibility scores from observed scores versus scores with imputation and weighting. The average intelligibility as a function of age was unchanged by the procedure and differences at the child-level were small. We interpret the similarity here as encouraging: The goal of this procedure was provide a coherent way to handle missing data, not dramatically change the results.