Additional Supplementary Material: Partial Analyses with 119 Participants (Outlier Candidate Removed)

1. Intelligibility (HI recordings) Results

As shown in Table 1, the selected random effects structure was Model 9, a random-intercept-only model. Likelihood ratio tests (LRTs) for the fixed effects (Table 2) indicated significant main effects of L1 and pre-test score. No significant issues were detected in any of the diagnostic tests conducted with the DHARMa package (v0.4.7; Hartig, 2024) for Model 11. The conditional R^2 and marginal R^2 were 0.590 and 0.084, respectively. Table 3 summarises the model output.

Table 1: *Random Effects Model Comparison for Intelligibility Scores (HI recordings)*

Sample size		Total observations = 5593 Subjects = 119; Items = 47									
Selection process	Model No.	Random effects		Model fit			Singular?	LRT test			
		Subject	Items	AIC	BIC	logLik		Reference model	df	χ^2	*p*
1. Building a maximal model											
	1	(1 | sub)	(1 + L1*Sound*Caption | item)	3597.1	3882.2	-1755.6	Yes	-	-	-	-
2. Building a ZCP model											
	2	(1 | sub)	(1 + L1*Sound*Caption || item)	3677.1	4227.4	-1755.6	Yes	-	-	-	-
3. Simplifying the ZCP model with LRT											
	3	(1 | sub)	(1 + L1:Sound + L1:Caption + Sound:Caption || item)	3602.7	3854.6	-1763.3	Yes	-	-	-	-
	4	(1 | sub)	(1 + L1:Sound + L1:Caption || item)	3582.7	3768.3	-1763.3	Yes	-	-	-	-
	5	(1 | sub)	(1 + Sound + L1:Sound || item)	3571.7	3710.9	-1764.9	Yes	-	-	-	-
	6	(1 | sub)	(1 + L1 + Sound + Caption || item)	3571.6	3684.3	-1768.8	Yes	-	-	-	-
	7	(1 | sub)	(1 + L1 + Sound || item)	3565.6	3658.4	-1768.8	Yes	-	-	-	-
	8	(1 | sub)	(1 + Sound || item)	3570.1	3643	-1774.1	Yes	-	-	-	-
	9	(1 | sub)	(1 | item)	3569.5	3622.5	-1776.7	No	-	-	-	-
	10	-	(1 | item)	3657.2	3703.6	-1821.6	No	9	1	89.67	< .001

Table 2: *Fixed Effects Model Comparisons for Intelligibility Scores (HI recordings)*

Sample size		Total observations = 5593 Subjects = 119; Items = 47							
Model name	Model No.	Fixed effects	Model fit			LRT Test			
			AIC	BIC	logLik	Reference model	df	χ^2	*p*
1. Interaction model									
	9	Sound + Caption + L1 + PreScore + Sound:Caption	3569.5	3622.5	-1776.7	-	-	-	-
2. No interaction model									
	11	Sound + Caption + L1 + PreScore	3567.5	3613.9	-1776.7	9	1	0.00	.971
3. No Sound model									
	12	Caption + L1 + PreScore	3565.5	3605.3	-1776.7	11	1	0.00	.953
4. No Caption model									
	13	Sound + L1 + PreScore	3566.5	3606.3	-1777.3	11	1	1.04	.309
5. No L1 model									
	14	Sound + Caption + PreScore	3571.6	3611.4	-1779.8	11	1	5.35	.013
6. No Pre-test model									
	16	Sound + Caption + L1							< .001

Table 3: *Results of Model 11 for Intelligibility Scores (HI recordings)*

	Fixed effects			Random effects	
				By subject	By items
Parameters	Estimate	*SE*	*z*	*SD*	*SD*
Intercept	2.86	0.32	8.94	0.68	1.87
Sound	-0.02	0.15	-0.13	-	-
Caption	-0.18	0.15	-1.16	-	-
L1	-0.41	0.15	-2.68	-	-
PreScore (covariate)	0.79	0.08	10.27	-	-

2. Intelligibility (LI recordings) Results

The initial random effects structure selection indicated that Model 7 was the optimal random effects structure (see Table 4).

Model 7: res ~ Sound*Caption + L1 + Pretest + (1 | sub) + (1 + L1 || item)

However, as noted in the main thesis text (p. 296), this model produced a singular fit warning when testing the main effects of L1. Therefore, the structure selection process was repeated, and Model 9, a random-intercept-only model, was selected.

LRTs for the main effects (Table 5) showed that caption, L1, and pre-test score had significant effects. However, the coefficient for caption was negative (Table 6), indicating a negative effect. Assessment of the no-interaction model (Model 12) revealed a few diagnostic issues in the full model and the subject-grouping model, but visual checks suggested these were minor: 1) the Kolmogorov–Smirnov test reached significance (p = .030); nevertheless, the QQ plot showed near-linear conformity with the 45° reference line in the full model; 2) the combined adjusted quantile test was significant, yet the residual-versus-predicted plot indicated that only the upper (0.75) quantile curve deviated slightly above the expected horizontal reference. The conditional and marginal R^2 were 0.606 and 0.069, respectively.

Table 4: *Random Effects Model Comparison for Intelligibility Scores (LI recordings)*

Sample size	Total observations = 5712 Subjects = 119; Items = 48									
Selection process	Model No.	Random effects		Model fit			Singular?	LRT test			
		Subject	Items	AIC	BIC	logLik		Reference model	df	χ^2	*p*
1. Building a maximal model											
	1	(1 | sub)	(1 + L1*Sound*Caption | item)	4974.5	5260.5	-2444.3	Yes	-	-	-	-
2. Building a ZCP model											
	2	(1 | sub)	(1 + L1*Sound*Caption || item)	5054.5	5606.5	-2444.3	Yes	-	-	-	-
3. Simplifying the ZCP model with LRT											
	3	(1 | sub)	(1 + L1:Sound + L1:Caption + Sound:Caption || item)	4973.3	5226.1	-2448.7	Yes	-	-	-	-
	4	(1 | sub)	(1 + L1:Sound + Sound:Caption || item)	4955.2	5141.4	-2449.6	Yes	-	-	-	-
	5	(1 | sub)	(1 + L1 + Sound:Caption || item)	4941.6	5081.3	-2449.8	Yes	-	-	-	-
	6	(1 | sub)	(1 + L1 + Sound + Caption || item)	Convergence problem			-	-	-	-	-
	7	(1 | sub)	(1 + L1 + Sound || item)	4935	5028.1	-2453.5	No	-	-	-	-
	8	(1 | sub)	(1 + L1 || item)	4929.6	5002.7	-2453.8	No	7	3	0.56	.905
	9	(1 | sub)	(1 | item)	4938.9	4992.2	-2461.5	No	8	3	15.36	.002
	11	-	(1 | item)	5094.5	5141.1	-2540.3	No	9	1	157.59	< .001
4. Comparing the selected model in LRT process with the model with correlation parameters											
	10	(1 | sub)	(1 + L1 | item)	4927.6	4994.1	-2453.8	No	8	1	0	.998

Table 5: *Fixed Effects Model Comparison for Intelligibility Scores (LI recordings)*

Sample size		Total observations = 5712 Subjects = 119; Items = 48							
Model name	Model No.	Fixed effects	Model fit			LRT Test			
			AIC	BIC	logLik	Reference model	df	χ^2	*p*
1. Interaction model									
	9	Sound + Caption + L1 + PreScore + Sound:Caption	4938.9	4992.2	-2461.5	-	-	-	-
2. No interaction model									
	12	Sound + Caption + L1 + PreScore	4939.1	4985.7	-2462.6	11	1	2.19	.139
3. No Sound model									
	13	Caption + L1 + PreScore	4939.2	4979.2	-2463.6	12	1	2.11	.146
4. No Caption model									
	14	Sound + L1 + PreScore	4941.9	4981.8	-2464.9	12	1	4.71	.030
5. No L1 model									
	15	Sound + Caption + PreScore	4948.6	4988.5	-2468.3	12	1	11.44	< .001
6. No Pre-test model									
	16	Sound + Caption + L1	4995.3	5035.3	-2491.7	12	1	58.21	< .001

Table 6: *Results of Model 12 for Intelligibility Scores (LI recordings)*

	Fixed effects			Random effects	
				By Subject	By Items
Parameters	Estimate	*SE*	*z*	*SD*	*SD*
Intercept	1.37	0.34	4.08	0.67	2.07
Sound	0.21	0.14	1.46	-	-
Caption	-0.32	0.14	-2.20	-	-
L1	-0.51	0.15	-3.47	-	-
PreScore (covariate)	0.63	0.07	8.61	-	-

3. Comprehensibility (HI recordings) Results

The selected random effects model was Model 9, a random-intercept-only model (see Table 7). LRTs for the fixed effects (Table 8) indicated that only the pre-test score reached statistical significance. Assessment of the no-interaction model (Model 11) revealed a few diagnostic issues in the full model, but visual checks suggested these were minor: 1) the KS test reached significance (p = .001); nevertheless, the QQ plot showed near-linear conformity with the 45° reference line; 2) although the combined adjusted quantile test was significant, deviations in the residual-versus-predicted plot were not substantial.

The conditional and marginal R^2 were 0.452 and 0.230, respectively. The model output is summarised in Table 9.

Table 7: *Random Effects Model Comparison for Comprehensibility Ratings (HI recordings)*

Sample size	Total observations = 1309 Subjects = 119; Items = 11									
Selection process	Model No.	Random effects		Model fit			Singular?	LRT test			
		Subject	Items	AIC	BIC	logLik		Reference model	df	χ^2	*p*
1. Building a maximal model											
	1	(1 | sub)	(1 + L1*Sound*Caption | item)	5157.5	5385.3	-2534.8	Yes	-	-	-	-
2. Building a ZCP model											
	2	(1 | sub)	(1 + L1*Sound*Caption || item)	5237.5	5672.4	-2534.8	Yes	-	-	-	-
3. Simplifying the ZCP model with LRT											
	3	(1 | sub)	(1 + L1:Sound + L1:Caption + Sound:Caption || item)	5154.6	5356.5	-2538.3	Yes	-	-	-	-
	4	(1 | sub)	(1 + L1:Caption + Sound:Caption || item)	5136.9	5287.0	-2539.4	Yes	-	-	-	-
	5	(1 | sub)	(1 + L1 + Sound:Caption || item)	5128.8	5242.6	-2542.4	Yes	-	-	-	-
	6	(1 | sub)	(1 + L1 + Sound + Caption || item)	5123.9	5217.1	-2543.9	Yes	-	-	-	-
	7	(1 | sub)	(1 + Sound + L1 || item)	5118.3	5196.0	-2544.1	Yes	-	-	-	-
	8	(1 | sub)	(1 + Sound || item)	5131.7	5193.8	-2553.9	No	-	-	-	-
	9	(1 | sub)	(1 | item)	5127.0	5173.6	-2554.5	No	8	3	1.33	.722
	10	-	(1 | item)	5176.6	5218.0	-2580.3	No	9	1	51.56	< .001

Table 8: *Fixed Effects Model Comparison for Comprehensibility Ratings (HI recordings)*

Sample size		Total observations = 1309 Subjects = 119; Items = 11							
Model name	Model No.	Fixed effects	Model fit			LRT Test			
			AIC	BIC	logLik	Reference model	df	χ^2	*p*
1. Interaction model									
	9	Sound + Caption + L1 + PreScore + Sound:Caption	5115.6	5162.1	-2548.8	-	-	-	-
2. No interaction model									
	11	Sound + Caption + L1 + PreScore	5115.4	5156.8	-2549.7	9	1	1.81	.179
3. No Sound model									
	12	Caption + L1 + PreScore	5113.4	5149.6	-2549.7	11	1	0.00	.961
4. No Caption model									
	13	Sound + L1 + PreScore	5113.5	5149.8	-2549.8	11	1	0.18	.667
5. No L1 model									
	14	Sound + Caption + PreScore	5113.9	5150.1	-2549.9	11	1	0.51	.475
6. No Pre-test model									
	16	Sound + Caption + L1							< .001

Table 9: *Results of Model 11 for Comprehensibility Ratings (HI recordings)*

	Fixed effects			Random effects	
				By subject	By items
Parameters	Estimate	*SE*	*t*	*SD*	*SD*
Intercept	5.63	0.29	19.34	0.58	0.84
Sound	0.01	0.14	0.05	-	-
Caption	-0.06	0.14	-0.42	-	-
L1	-0.10	0.14	-0.70	-	-
PreScore (covariate)	1.05	0.07	14.63	-	-

4. Comprehensibility (LI recordings) Results

Based on LRTs, Model 9, a random-intercept-only model, was selected as the optimal random effects model (see Table 10). As shown in Table 11, LRTs for the main effects indicated that exposure accent and pre-test score reached statistical significance. Assessment of the no-interaction model (Model 11) showed that the outlier test was significant (p = .042). The conditional and marginal R^2 were 0.493 and 0.215, respectively. The model output is summarised in Table 12.

Table 10: *Random Effects Model Comparison for Comprehensibility Ratings (LI recordings)*

Sample size		Total observations = 1309 Subjects = 119; Items = 11									
Selection process	Model No.	Random effects		Model fit			Singular?	LRT test			
		Subject	Items	AIC	BIC	logLik		Reference model	df	χ^2	*p*
1. Building a maximal model											
	1	(1 | sub)	(1 + L1*Sound*Caption | item)	5147.5	5375.3	-2529.7	Yes	-	-	-	-
2. Building a ZCP model											
	2	(1 | sub)	(1 + L1*Sound*Caption || item)	5228.1	5663.0	-2530.0	Yes	-	-	-	-
3. Simplifying the ZCP model with LRT											
	3	(1 | sub)	(1 + L1:Sound + L1:Caption + Sound:Caption || item)	5143.9	5345.8	-2532.9	Yes	-	-	-	-
	4	(1 | sub)	(1 + L1:Sound + L1:Caption || item)	5126.7	5276.8	-2534.3	Yes	-	-	-	-
	5	(1 | sub)	(1 + Sound + L1:Caption || item)	5114.7	5228.6	-2535.3	Yes	-	-	-	-
	6	(1 | sub)	(1 + L1 + Sound + Caption || item)	5109.3	5202.5	-2536.7	Yes	-	-	-	-
	7	(1 | sub)	(1 + L1 + Caption || item)	5103.4	5181.1	-2536.7	Yes	-	-	-	-
	8	(1 | sub)	(1 + Caption || item)	5101.6	5163.7	-2538.8	Yes	-	-	-	-
	9	(1 | sub)	(1 | item)	5095.6	5142.2	-2538.8	No	-	-	-	-
	10	-	(1 | item)	5155.6	5197.0	-2569.8	No	9	1	61.92	< .001

Table 11: *Fixed Effects Model Comparison for Comprehensibility Ratings (LI recordings)*

Sample size		Total observations = 1309 Subjects = 119; Items = 11							
Model name	Model No.	Fixed effects	Model fit			LRT Test			
			AIC	BIC	logLik	Reference model	df	χ^2	*p*
1. Interaction model									
	9	Sound + Caption + L1 + PreScore + Sound:Caption	5084.6	5131.2	-2533.3	-	-	-	-
2. No interaction model									
	11	Sound + Caption + L1 + PreScore	5084.8	5126.2	-2534.4	9	1	2.13	.144
3. No Sound model									
	12	Caption + L1 + PreScore	5086.8	5123	-2536.4	11	1	4.00	.046
4. No Caption model									
	13	Sound + L1 + PreScore	5084	5120.2	-2535	11	1	1.20	.273
5. No L1 model									
	14	Sound + Caption + PreScore	5086.2	5122.4	-2536.1	11	1	3.44	.064
6. No Pre-test model									
	16	Sound + Caption + L1							< .001

Table 12: R*esults of Model 11 for Comprehensibility Ratings (LI recordings)*

	Fixed effects			Random effects	
				By Subject	By Items
Parameters	Estimate	*SE*	*t*	*SD*	*SD*
Intercept	4.52	0.33	13.50	0.61	1.00
Sound	0.28	0.14	1.98	-	-
Caption	-0.15	0.14	-1.08	-	-
L1	-0.26	0.14	-1.84	-	-
PreScore (covariate)	1.01	0.07	14.00	-	-